Multivariate Regression Analysis of Academic Achievement: Identifying Key Determinants of Student Performance Outcomes

S. Hariprasad[1], D. Giri[2], K. Sreenivasulu[3], M. Pedda Reddeppa Reddy[4], K. Murali[5**]

1. Associate Professor, Dept. of Statistics, PVKN Govt. College (A), Chittoor.
2. Principal, D.K. Govt. College for Women(A), Nellore.
3. Lecturer in Statistics, Sri Padmavathi Women's Degree and PG College, TTD(A), Tirupati.
4. Associate Professor, Department of Statistics, S V Arts college, TTD(A), Tirupati.
5. Department of Statistics, S V University, Tirupati.

[**]Corresponding Author: **dr.murali4stat@gmail.com**

Abstract

This study presents a comprehensive multivariate regression analysis of academic achievement determinants using the xAPI Learning Analytics Education dataset, comprising 480 students enrolled in the Kalboard 360 Learning Management System. The research employs a hierarchical ordinary least squares (OLS) regression framework across four progressive model variants — Behavioral, Behavioral + Parental, Behavioral + Parental + Demographic, and Full Model — to quantify the incremental explanatory power of each predictor category on student performance outcomes measured on an ordinal three-class scale (Low = 1, Medium = 2, High = 3).

The full multivariate regression model achieves $R^2 = 0.6849$ and Adjusted $R^2 = 0.6660$ ($F = 70.43$, $p < 0.001$), explaining 68.5% of variance in student academic performance — a strong fit for an educational social science dataset. Hierarchical regression reveals that behavioral engagement predictors alone account for 52.3% of variance (Model 1), with parental involvement adding a significant 4.8 percentage points ($\Delta R^2 = 0.048$, Model 2), demographic variables contributing a further 1.4 points (Model 3), and student absence contributing the largest single increment of 9.1 points ($\Delta R^2 = 0.091$, Model 4). Ten-fold cross-validation confirms generalizability with CV $R^2 = 0.657 \pm 0.107$ and CV RMSE $= 0.425 \pm 0.059$.

Standardized coefficient analysis identifies Student Absence ($\beta^* = -0.380$) and Visited Resources ($\beta^* = +0.202$) as the two most influential determinants, followed by Raised Hands ($\beta^* = +0.181$), Parent

Answering Survey ($\beta^* = +0.137$), and Parent-Guardian Relation ($\beta^* = +0.114$). All regression assumptions are satisfactorily met: residuals are normally distributed (Shapiro-Wilk $W = 0.997$, $p = 0.808$), homoscedastic (Breusch-Pagan $p = 0.182$), non-autocorrelated (Durbin-Watson = 2.105), and all VIF values are below 2.5 — confirming the absence of problematic multicollinearity. Chi-square tests confirm that both Student Absence ($\chi^2 = 225.20$, $p < 0.001$) and Parental Survey Participation ($\chi^2 = 95.36$, $p < 0.001$) are highly significant categorical determinants of academic class.

Keywords: Multivariate Regression, Academic Achievement, Student Performance, xAPI Learning Analytics, Hierarchical Regression, Behavioral Engagement, Parental Involvement, Student Absence, Educational Data Mining, Learning Management System, OLS Regression, Determinants of Learning

1. Introduction

1.1 Background and Motivation

Academic achievement is the cornerstone outcome of educational systems worldwide, yet it emerges from an intricate interplay of behavioral, familial, demographic, and institutional factors that no single predictor can adequately explain in isolation. The exponential growth of digital learning environments — particularly Learning Management Systems (LMS) — has generated unprecedented volumes of student interaction data, creating new opportunities to empirically model the determinants of academic performance with greater precision than was previously possible from survey-based or administrative records alone.

The xAPI (Experience API) specification, also known as Tin Can API, enables the systematic capture of granular learner interactions with digital content — including how frequently students raise hands in virtual classrooms, how often they visit learning resources, how many announcements they view, and the depth of their participation in discussion forums. These behavioral traces, combined with parental engagement indicators and demographic attributes, constitute a rich multivariate predictor space for modeling academic outcomes. Understanding which of these factors most strongly and independently predict student performance has direct implications for early warning system design, targeted pedagogical intervention, and educational resource allocation.

Despite the proliferation of classification-oriented machine learning studies on student performance datasets, the regression-analytical tradition — which provides interpretable coefficient estimates, statistical inference on individual predictors, and model diagnostic validation — remains underrepresented in the educational data mining literature. Multivariate ordinary least squares (OLS) regression, and its hierarchical extension, offer unique advantages: the ability to quantify the marginal effect of each predictor while controlling for all others, to test the incremental explanatory power of theoretically meaningful predictor blocks, and to generate policy-actionable standardized effect sizes (β*) that facilitate direct comparisons across predictors measured on different scales.

1.2 Research Objectives

This study is structured around five primary research objectives:

1. To characterize the distribution of student academic performance across Low, Medium, and High achievement classes and examine class-conditional differences in behavioral engagement, parental involvement, and demographic attributes.
2. To construct and compare four hierarchical regression models — progressively incorporating behavioral, parental, demographic, and absence predictors — and quantify the incremental R^2 contribution of each predictor block.
3. To identify and rank the key determinants of student performance through standardized regression coefficients (β*), Pearson correlations, and Kruskal-Wallis hypothesis tests.
4. To validate regression model assumptions through residual normality (Shapiro-Wilk, Jarque-Bera), homoscedasticity (Breusch-Pagan), autocorrelation (Durbin-Watson), and multicollinearity (VIF) diagnostics.
5. To assess the generalizability of the full regression model through 10-fold cross-validation and provide evidence-based recommendations for educational policy and early intervention system design.

1.3 Study Context: Kalboard 360 LMS

The xAPI-Edu-Data dataset originates from the Kalboard 360 Learning Management System, a cloud-based educational platform used by students across Kuwait, Jordan, Egypt, and other Middle Eastern and North African countries. The dataset captures interactions of 480 students across 12 academic subjects (including IT, Mathematics, Arabic, Science, English, Biology, and Chemistry) across lower, middle, and high school educational stages. This multi-subject, multi-nationality, multi-stage composition provides a diverse and generalizable analytical context for identifying robust predictors of academic achievement.

2. Literature Review

2.1 Determinants of Academic Performance: Theoretical Framework

The theoretical foundations of academic achievement determinants draw from three major psychological and sociological frameworks. Bandura's (1977) Social Cognitive Theory emphasizes behavioral engagement and self-efficacy as primary academic performance drivers — students who actively participate (raising hands, visiting resources, engaging in discussions) develop stronger metacognitive regulation and deeper content mastery. Bronfenbrenner's (1979) Ecological Systems Theory highlights the nested influence of family and school environments — parental involvement, guardian-teacher communication, and institutional satisfaction represent the mesosystem linkages that amplify or attenuate a student's individual academic effort. Coleman's (1966) educational sociology research established the profound influence of socioeconomic and family factors on academic outcomes, a finding replicated in hundreds of subsequent studies.

Empirical regression analyses of academic performance consistently identify three dominant predictor clusters: academic engagement behaviors (attendance, participation, study time), parental and home environment factors (parental education, involvement, socioeconomic status), and demographic characteristics (gender, age, school level). The relative magnitude of these clusters varies by dataset and educational context, but their directional relationships — engagement positively predicts performance; absence negatively predicts performance; parental involvement positively predicts performance — are among the most replicated findings in educational research.

2.2 Behavioral Engagement and Academic Outcomes

The relationship between active learning behaviors and academic achievement is well-established in both experimental and observational research. Meta-analyses of classroom participation studies consistently find participation-performance correlations in the range $r = 0.35–0.55$, consistent with the $r = 0.646$ for Raised Hands and $r = 0.677$ for Visited Resources observed in this study. Tinto's (1987) student integration model posits that academic and social integration — operationalized in digital environments through resource visits, discussion engagement, and announcement monitoring — are the strongest proximate determinants of academic success and persistence.

The role of digital resource utilization as an academic performance predictor has gained prominence with the expansion of LMS-based education. Research on LMS interaction data consistently finds that resource access frequency is among the strongest behavioral predictors of academic outcomes, with students in higher performance bands accessing learning materials 3–5 times more frequently than lower-performing peers — a pattern evident in the xAPI data, where High-class students average 78.75 resource visits compared to 18.32 for Low-class students.

2.3 Parental Involvement

Parental involvement is among the most consistently replicated determinants of K-12 academic achievement in the international literature. Jeynes' (2007) meta-analysis of 52 studies found that parental involvement has an overall effect size equivalent to moving a student from the 50th to the 60th percentile. The mechanisms through which parental involvement operates include direct academic support (assistance with homework, monitoring of study habits), motivational scaffolding (communicating academic expectations), and school-family communication (attending parent-teacher meetings, responding to school surveys). In the xAPI dataset, the parental survey participation variable captures this school-family communication channel: 56.2% of parents participated in the survey, and survey-participating parents' children achieve High classification at a rate of 42.2% compared to only 13.3% for non-participating parents.

2.4 Student Absence and Academic Risk

Chronic absenteeism — conventionally defined as missing 10% or more of school days — is one of the strongest and most replicated predictors of academic underachievement, grade retention, and dropout. Research by Chang and Romero (2008) found that chronic absence in early grades predicts third-grade reading proficiency, and its effects compound through subsequent grade levels. In the xAPI dataset, the binary absence variable (Above-7 days absent vs. Under-7 days) demonstrates the most powerful individual association with academic class: 91.3% of Low-class students are absent more than 7 days, compared to only 2.8% of High-class students — a 32-fold difference that makes absence the single most discriminating predictor in the dataset.

2.5 Multivariate Regression in Educational Research

Multivariate OLS regression has been the workhorse method of educational outcome research for decades. Hierarchical (blockwise) regression — in which theoretically meaningful predictor groups are entered sequentially and ΔR^2 is assessed at each stage — allows researchers to partition explained variance among predictor categories and test the incremental contribution of each block after controlling for prior blocks. This approach aligns with theoretical frameworks that posit behavioral factors as more proximate (and therefore primary) determinants of academic outcomes, with family and demographic factors as more distal (secondary) influences that operate partly through their effects on engagement.

3. Data Description

3.1 Dataset Overview

The xAPI Learning Analytics Education dataset contains 480 student records collected from the Kalboard 360 Learning Management System. The dataset is fully complete — no missing values across any of the 17 variables — a significant data quality advantage relative to most educational datasets. Students are drawn from lower level, middle school, and high school stages, spanning 12 academic subjects and representing 14 nationalities primarily from the Middle East and North Africa region.

Variable	Type	Description	Range / Categories
raisedhands	Continuous	Times student raised hand in virtual class (behavioral engagement)	0 – 100
VisITedResources	Continuous	Number of times student visited course content/materials	0 – 99
AnnouncementsView	Continuous	Number of times student viewed course announcements	0 – 98
Discussion	Continuous	Number of times student participated in discussion groups	1 – 99
StudentAbsenceDays	Binary	Absence level: Under-7 days or Above-7 days	Under-7 / Above-7
ParentAnsweringSurvey	Binary	Whether parent responded to school satisfaction survey	Yes / No
ParentschoolSatisfaction	Binary	Parent-reported school satisfaction rating	Good / Bad
Relation	Binary	Primary guardian of student (father or mother)	Father / Mum
gender	Binary	Student gender	M / F
StageID	Ordinal	Educational stage of student	Lower / Middle / High School
Semester	Binary	Academic semester of the record	First / Second
NationalITy	Nominal	Student nationality (14 countries)	KW, Jordan, Egypt, etc.
Topic	Nominal	Academic subject studied (12 subjects)	IT, Math, Arabic, etc.
Class (Outcome)	Ordinal	Academic performance class: Low / Medium / High	L=1, M=2, H=3

Table 1: Variable Descriptions — xAPI-Edu-Data (N = 480)

3.2 Outcome Variable: Academic Performance Classification

The outcome variable Class represents each student's academic performance tier based on their cumulative assessment results within the LMS: Low (L) corresponds to scores below 69 (score range 0–69), Medium (M) to scores 70–89, and High (H) to scores 90–100. For regression analysis, this ordinal variable is numerically encoded as L = 1, M = 2, H = 3, treating the ordinal scale as approximately interval — a standard practice in educational regression research for Likert-type and graded outcome variables.

Class	Label	Score Range	Count	Proportion	Cumulative %
H	High Achievers	90 – 100	142	29.6%	29.6%
M	Medium Achievers	70 – 89	211	44.0%	73.5%
L	Low Achievers	0 – 69	127	26.5%	100.0%
Total	—	0 – 100	480	100.0%	—

Table 2: Academic Performance Class Distribution (N = 480)

The class distribution is broadly balanced — Medium achievers constitute the plurality (44.0%), with High and Low achievers comprising 29.6% and 26.5% respectively. This near-symmetric, unimodal distribution around the Medium class provides a favorable environment for regression modeling, minimizing class imbalance concerns.

3.3 Descriptive Statistics

Feature	Mean	Std Dev	Min	Median	Max	Class L Mean	Class M Mean	Class H Mean
raisedhands	46.78	30.78	0	50.0	100	16.9	48.9	70.3
VisITedResources	54.80	33.08	0	65.0	99	18.3	60.6	78.7
AnnouncementsView	37.92	26.61	0	33.0	98	15.6	41.0	53.4
Discussion	43.28	27.64	1	39.0	99	30.8	43.8	53.7

Table 3: Descriptive Statistics — Behavioral Engagement Features by Academic Class

The between-class differences in behavioral engagement are large and monotonically ordered. High achievers raise hands an average of 70.3 times — 4.16× more than Low achievers (16.9) and 1.44× more than Medium achievers (48.9). The gradient is even steeper for resource visits: High-class students average 78.75 visits versus 18.32 for Low-class students, a 4.30× ratio that represents the largest relative behavioral gap in the dataset. These large effect sizes across all four behavioral indicators foreshadow their strong predictive power in the regression models.

3.4 Categorical Variable Summary

Variable	Category	N (%)	% Low	% Medium	% High	Chi² (vs Class)	p-value
Student Absence	Above-7 days	191 (39.8%)	60.7%	37.2%	2.1%	225.20	$p < 0.001$
Student Absence	Under-7 days	289 (60.2%)	3.8%	48.4%	47.8%	—	—
Parent Survey	Yes (Participated)	270 (56.3%)	10.4%	47.4%	42.2%	95.36	$p < 0.001$
Parent Survey	No (Did not)	210 (43.8%)	47.1%	39.5%	13.3%	—	—

Variable	Category	N (%)	% Low	% Medium	% High	Chi^2 (vs Class)	p-value
Gender	Female	175 (36.5%)	13.7%	43.4%	42.9%	—	—
Gender	Male	305 (63.5%)	33.8%	44.3%	22.0%	—	—
Stage	High School	—	24.2%	42.4%	33.3%	—	—
Stage	Middle School	—	21.8%	47.6%	30.6%	—	—
Stage	Lower Level	—	32.7%	39.7%	27.6%	—	—

Table 4: Categorical Variable Distributions by Academic Class with Chi-Square Tests

The cross-tabulation analysis reveals striking distributional patterns. Student absence is the most powerful categorical discriminator: among students absent more than 7 days, 60.7% fall in the Low class and only 2.1% achieve High classification — a near-polar distribution. Conversely, among students with under 7 days absence, 47.8% achieve High and only 3.8% fall in Low. The chi-square statistic of 225.20 (df = 2, $p < 0.001$) for this relationship is the largest in the dataset, confirming absence as the dominant categorical determinant of academic class. Gender also shows a meaningful pattern: female students achieve High classification at 42.9% compared to 22.0% for male students — a 20.9 percentage point gap.

4. Methodology

4.1 Variable Encoding

All categorical predictors were numerically encoded for regression analysis using logical binary mappings that preserve ordinal direction. Absent Above-7 was encoded as 1 (higher absence = higher score = negative for performance), Parent Answering Survey Yes = 1, Parent School Satisfaction Good = 1, gender Female = 1, Relation Mum = 1, and Semester Second = 1. Educational stage was encoded ordinally as Lower Level = 1, Middle School = 2, High School = 3. No one-hot encoding was required as all categorical variables are binary (two levels) or logically ordinal (educational stage).

4.2 Hierarchical Multivariate OLS Regression

The study employs hierarchical (blockwise) ordinary least squares regression with four theoretically motivated predictor blocks entered sequentially:

Model	Predictor Block Added	Variables Included	No. of Predictors
Model 1 (Behavioral)	Behavioral Engagement	raisedhands, VisITedResources, AnnouncementsView, Discussion	4
Model 2 (+Parental)	Parental Involvement	+ ParentAnsweringSurvey, ParentschoolSatisfaction, Relation	7
Model 3 (+Demographic)	Student Demographics	+ gender, StageID, Semester	10
Model 4 (Full)	Absence Behavior	+ StudentAbsenceDays	11

Table 5: Hierarchical Regression Model Blocks

This block-entry strategy reflects the theoretical hierarchy of academic performance determinants: behavioral engagement is the most proximate (and therefore primary) predictor; parental factors are distal influences that shape behavioral dispositions; demographic characteristics are background controls; and absence, while behavioral in nature, is treated as a distinct block due to its known dominant effect size in educational research.

4.3 The Full Regression Equation

The complete multivariate regression equation for Model 4 (Full Model) is:

$$\text{ClassNum} = \beta_0 + \beta_1 \cdot \text{RaisedHands} + \beta_2 \cdot \text{VisitedResources} + \beta_3 \cdot \text{AnnouncementsView}$$

$$+ \beta_4 \cdot \text{Discussion} + \beta_5 \cdot \text{ParentSurvey} + \beta_6 \cdot \text{ParentSatisfaction}$$

$$+ \beta_7 \cdot \text{Relation} + \beta_8 \cdot \text{Gender} + \beta_9 \cdot \text{StageID} + \beta_{10} \cdot \text{Semester}$$

$$+ \beta_{11} \cdot \text{AbsenceDays} + \varepsilon$$

where ClassNum $\in \{1, 2, 3\}$ is the ordinal outcome, β_0 is the intercept, β_1–β_{11} are the partial regression coefficients, and $\varepsilon \sim N(0, \sigma^2)$ is the error term.

4.4 Standardized Coefficients

To enable direct comparison of predictor effects across variables measured on different scales, standardized beta coefficients (β^*) are computed as:

$$\beta^*_k = \beta_k \cdot (SD_{xk} / SD_Y)$$

where SD_{xk} is the standard deviation of predictor k, SD_Y is the standard deviation of the outcome, and β_k is the unstandardized regression coefficient. β^* represents the expected change in the outcome in standard deviation units for a one-standard-deviation change in the predictor, holding all other predictors constant.

4.5 Regression Diagnostics

Assumption	Test Applied	Null Hypothesis	Decision Rule
Normality of Residuals	Shapiro-Wilk, Jarque-Bera	Residuals are normally distributed	Fail to reject H_0 ($p > 0.05$)
Homoscedasticity	Breusch-Pagan LM test	Constant residual variance	Fail to reject H_0 ($p > 0.05$)
No Autocorrelation	Durbin-Watson statistic	No serial correlation in residuals	DW ≈ 2.0 (range 1.5–2.5)
No Multicollinearity	Variance Inflation Factor (VIF)	Predictors not linearly dependent	VIF < 5 acceptable; < 2.5 ideal
Model Significance	F-test (ANOVA)	All coefficients jointly = 0	Reject H_0 ($p < 0.05$)
Individual Predictors	t-tests on coefficients	$\beta_k = 0$ (no effect)	Reject H_0 ($p < 0.05$)

Table 6: Regression Diagnostic Tests and Decision Rules

4.6 Model Validation

Model generalizability is assessed through 10-fold cross-validation on the full dataset, partitioning observations into 10 equal folds, training on 9 folds, and evaluating on the held-out fold — repeating this process 10 times and averaging performance metrics. The primary cross-validation metrics are R^2 (coefficient of determination) and RMSE (root mean square error). An 80/20 stratified hold-out split (384 training, 96 test observations) is additionally used for final model evaluation and out-of-sample performance reporting.

5. Empirical Results

5.1 Bivariate Correlation Analysis

Predictor	Pearson r with ClassNum	Significance	Direction	Interpretation
VisITedResources	+0.677	$p < 0.001$ ***	Positive	Strongest positive predictor — resource engagement
StudentAbsenceDays (encoded)	−0.671	$p < 0.001$ ***	Negative	Absence strongly reduces performance class
raisedhands	+0.646	$p < 0.001$ ***	Positive	Active classroom participation boosts achievement
AnnouncementsView	+0.527	$p < 0.001$ ***	Positive	Moderate — staying informed aids performance

Predictor	Pearson r with ClassNum	Significance	Direction	Interpretation
ParentAnsweringSurvey	+0.436	p < 0.001 ***	Positive	Parent engagement strongly predicts higher class
Relation (Mum=1)	+0.401	p < 0.001 ***	Positive	Mother-primary-guardian linked to higher achievement
ParentschoolSatisfaction	+0.376	p < 0.001 ***	Positive	School satisfaction reflects positive environment
Discussion	+0.308	p < 0.001 ***	Positive	Collaborative learning moderately predicts success
gender (Female=1)	+0.264	p < 0.001 ***	Positive	Female students achieve higher on average
Semester (Second=1)	+0.126	p = 0.006 **	Positive	Weak — slight second-semester advantage
StageID (Higher=3)	+0.084	p = 0.066 ns	Positive	Not significant — stage has minimal linear effect

*Table 7: Pearson Correlation Coefficients — All Predictors vs. Academic Performance Class (*** p<0.001, ** p<0.01)*

The bivariate correlation analysis reveals a clear hierarchy of univariate predictor importance. The top three correlates are Visited Resources ($r = +0.677$), Student Absence ($r = -0.671$, where higher encoding = more absence = lower performance), and Raised Hands ($r = +0.646$), forming a cluster of dominant behavioral predictors with near-equal explanatory strength. These are followed by Announcements View ($r = +0.527$) and three parental variables ($r = +0.376$ to $+0.436$), which form a secondary cluster of moderate predictors. Gender ($r = +0.264$, female = higher achievement) and Semester ($r = +0.126$) round out the significant predictors, while Stage shows no statistically significant linear correlation ($p = 0.066$).

5.2 Kruskal-Wallis Non-Parametric Tests

Behavioral Feature	H Statistic	p-value	Significance	Conclusion
VisITedResources	222.10	p < 0.000001	***	Highly significant class differences in resource visits
raisedhands	207.82	p < 0.000001	***	Highly significant class differences in participation
AnnouncementsView	154.85	p < 0.000001	***	Highly significant class differences in announcement views
Discussion	49.65	p < 0.000001	***	Highly significant — discussion participation varies by class

Table 8: Kruskal-Wallis Tests — Behavioral Engagement Features Across Academic Performance Classes

Kruskal-Wallis tests confirm statistically significant differences in the distribution of all four behavioral engagement features across the three academic performance classes (all $p < 0.000001$). The Visited Resources H-statistic of 222.10 is the highest, reflecting the largest between-class distributional differences — consistent with the High-class students visiting resources 4.3× more frequently than Low-class students. These non-parametric results validate the correlation findings without assuming normality within each class.

5.3 Hierarchical Regression Results

Model	Predictors	k	R^2	Adj-R^2	ΔR^2	F-statistic	p-value	RMSE (test)	MAE (test)
Model 1 — Behavioral	RH, VR, AV, DI	4	0.5230	0.5180	0.5230	103.91	< 0.001	0.4754	0.4021
Model 2 — +Parental	+Survey, Satisf., Relation	7	0.5712	0.5632	+0.0482	71.54	< 0.001	0.4631	0.3957
Model 3 — +Demographic	+Gender, Stage, Semester	10	0.5851	0.5740	+0.0139	52.60	< 0.001	0.4634	0.3914
Model 4 — Full Model ★	+StudentAbsence	11	0.6756	0.6660	+0.0905	70.43	< 0.001	0.3743	0.3038

Table 9: Hierarchical Regression Results — Four Model Variants (Train N=384, Test N=96) [★ = Selected Final Model]

The hierarchical regression results provide a clear and theoretically coherent picture of how variance in student academic performance accumulates across predictor blocks. Model 1, containing only the four behavioral engagement predictors, already accounts for a substantial 52.3% of variance in academic performance class (Adj-R^2 = 0.518, F = 103.91, $p < 0.001$) — confirming that in-platform learning behaviors are the single most powerful category of academic performance determinants in this dataset.

The addition of parental involvement variables in Model 2 produces a statistically meaningful increment of 4.82 percentage points ($\Delta R^2 = 0.048$, total $R^2 = 0.571$), confirming the independent contribution of parental engagement even after controlling for behavioral engagement. This incremental gain is practically significant: it indicates that family involvement explains approximately one-tenth as much variance as direct behavioral engagement, independently of how engaged the student is themselves.

Model 3's demographic predictors (gender, stage, semester) add a modest 1.39 percentage points ($\Delta R^2 = 0.014$, total $R^2 = 0.585$). While statistically meaningful at the level of the gender coefficient ($r = 0.264$, $p <$

0.001), stage and semester contribute minimally — suggesting that academic performance determinants are broadly consistent across school stages and semesters in this dataset. The largest single increment comes from adding Student Absence in Model 4 ($\Delta R^2 = 0.091$, total $R^2 = 0.676$) — a 9.05 percentage point jump that elevates the full model to Adj-$R^2 = 0.666$ and reduces test RMSE from 0.463 to 0.374, confirming absence as the dominant categorical determinant even after all behavioral and family predictors are controlled.

5.4 Full Model Regression Coefficients

Predictor	Unstd. Coef (β)	Std. Coef (β*)	Direction	Rank by |β*|
StudentAbsenceDays	−0.5797	−0.3798	↓ Negative	#1 — Dominant predictor
VisITedResources	+0.00457	+0.2023	↑ Positive	#2 — Resource engagement
raisedhands	+0.00440	+0.1810	↑ Positive	#3 — Active participation
ParentAnsweringSurvey	+0.2069	+0.1374	↑ Positive	#4 — Parental engagement
Relation (Mum=1)	+0.1730	+0.1139	↑ Positive	#5 — Guardian type
gender (Female=1)	+0.1194	+0.0769	↑ Positive	#6 — Gender advantage
AnnouncementsView	+0.00138	+0.0489	↑ Positive	#7 — Moderate behavioral
Discussion	+0.00143	+0.0529	↑ Positive	#8 — Collaborative learning
ParentschoolSatisfaction	+0.0457	+0.0298	↑ Positive	#9 — School environment
StageID	−0.0376	−0.0304	↓ Negative	#10 — Minimal stage effect
Semester	+0.00256	+0.0017	↑ Positive	#11 — Negligible
Intercept (β_0)	1.4938	—	—	—

Table 10: Full Model (Model 4) Regression Coefficients — Unstandardized and Standardized (N = 480)

The standardized coefficient analysis provides the definitive ranking of academic performance determinants in the full multivariate model, controlling for all other predictors simultaneously. Student Absence dominates with $\beta^* = -0.380$ — being absent more than 7 days reduces academic performance class by 0.38 standard deviations, holding all other factors constant. This represents the largest independent effect in the model, confirming that no amount of resource visits or parental engagement can fully compensate for chronic absenteeism.

Visited Resources ($\beta^* = +0.202$) and Raised Hands ($\beta^* = +0.181$) rank second and third — nearly equal in magnitude and both representing active in-platform engagement behaviors. Parent Answering Survey ($\beta^* = +0.137$) ranks fourth, indicating that parental communication with the school retains significant independent predictive value even after controlling for the student's own behavioral engagement. The guardian type variable Relation (β* = +0.114) — reflecting whether the primary guardian is the mother — ranks fifth, a finding consistent with international research on maternal primary guardianship and academic outcomes.

Gender (β* = +0.077, female positive) and Discussion (β* = +0.053) round out the moderate-effect predictors. Announcements View, Parent Satisfaction, Stage, and Semester all contribute β* values below 0.05, indicating minimal independent effects in the full multivariate context — though their univariate correlations are larger, reflecting suppression by the more powerful predictors that absorb their explanatory variance.

5.5 Cross-Validation Results

Metric	Value	Standard Deviation	95% CI (approx.)	Interpretation
10-Fold CV R^2	0.6568	±0.1065	[0.445, 0.869]	Strong — 65.7% variance explained on unseen data
10-Fold CV RMSE	0.4248	±0.0593	[0.306, 0.543]	Average prediction error < 0.5 class units
Test Set R^2	0.6756	—	—	Consistent with CV estimate
Test Set RMSE	0.3743	—	—	Below 0.5 — strong out-of-sample performance
Test Set MAE	0.3038	—	—	Mean absolute error < 0.31 class units

Table 11: 10-Fold Cross-Validation and Test Set Performance — Full Regression Model

The 10-fold cross-validation confirms that the full model generalizes well to unseen data, with mean CV R^2 = 0.657 — only 1.9 percentage points below the training R^2 of 0.676, indicating minimal overfitting. The CV RMSE of 0.425 (±0.059) means the model's predictions are on average less than half a class unit away from the true class. The relatively high standard deviation of CV R^2 (±0.107) reflects natural variation across 10-fold partitions of a 480-observation dataset, but the lower bound of approximately 0.44 still represents a reasonable model for fold-level variation.

6. Regression Diagnostic Analysis

6.1 Multicollinearity Diagnostics (VIF)

Predictor	VIF Value	Tolerance (1/VIF)	Assessment
raisedhands	2.478	0.404	Acceptable — moderate correlation with other behavioral vars
VisITedResources	2.432	0.411	Acceptable — correlated with raisedhands (both behavioral)
AnnouncementsView	2.321	0.431	Acceptable — shares behavioral engagement space
ParentAnsweringSurvey_n	1.628	0.615	Low — parental variables mostly independent
ParentschoolSatisfaction_n	1.577	0.634	Low — acceptable collinearity
Discussion	1.336	0.748	Low — most independent behavioral variable
Relation_n	1.310	0.763	Low — guardian variable independent of behavioral
Semester_n	1.137	0.880	Very low — no collinearity concern
gender_n	1.126	0.888	Very low — gender largely orthogonal
StudentAbsenceDays_n	1.433	0.698	Low — absence moderately correlated with engagement
StageID_n	1.078	0.928	Very low — stage level independent

Table 12: Variance Inflation Factor (VIF) Diagnostics — Full Regression Model

All eleven predictors exhibit VIF values below 2.5 — well within the conventional threshold of 5 (acceptable) and the conservative threshold of 10 (problematic). The three behavioral predictors (Raised Hands, Visited Resources, Announcements View) show the highest VIFs (2.3–2.5), reflecting their natural intercorrelation as jointly produced by the same underlying student engagement disposition. These modest VIFs indicate that multicollinearity is not a concern in this model — coefficient estimates are stable and interpretable as independent partial effects.

6.2 Residual Normality Tests

Test	Statistic	p-value	Conclusion
Shapiro-Wilk Normality Test	W = 0.9974	p = 0.8084	PASS — residuals are normally distributed
Jarque-Bera Normality Test	stat = 0.5872	p = 0.7456	PASS — residuals are normally distributed
Residual Mean	0.000000	—	PASS — zero-mean residuals (unbiased model)
Residual Std Dev	0.4312	—	Low dispersion — approximately half a class unit error

Table 13: Residual Normality Diagnostics — Full Model Training Residuals (n = 384)

Both the Shapiro-Wilk test ($W = 0.997$, $p = 0.808$) and the Jarque-Bera test (stat = 0.587, $p = 0.746$) strongly fail to reject the null hypothesis of normally distributed residuals — the gold standard diagnostic outcome for linear regression. This result is notable: with an ordinal outcome variable (ClassNum $\in \{1,2,3\}$), residual normality is not guaranteed, yet the central limit theorem operating over 384 training observations produces well-behaved Gaussian residuals. This validates the inference-based interpretation of F-statistics, t-tests, and confidence intervals reported for all model variants.

6.3 Homoscedasticity and Autocorrelation

Test	Statistic	p-value	Conclusion
Breusch-Pagan LM Test (Homoscedasticity)	stat = 15.019	p = 0.1816	PASS — residual variance is constant (homoscedastic)
Durbin-Watson Statistic (Autocorrelation)	DW = 2.1051	—	PASS — no serial autocorrelation (ideal range 1.5–2.5)

Table 14: Homoscedasticity and Autocorrelation Diagnostics

The Breusch-Pagan test ($p = 0.182$) fails to reject homoscedasticity — residual variance does not systematically increase or decrease as a function of fitted values. This confirms that OLS standard errors are valid and hypothesis tests are correctly sized. The Durbin-Watson statistic of 2.105 falls squarely in the no-autocorrelation zone (1.5–2.5), confirming that residuals are not serially correlated across the training observations.

6.4 Complete Diagnostic Summary

Assumption	Test	Result	Status
Residual Normality	Shapiro-Wilk	W=0.997, p=0.808 > 0.05	✓ PASS
Residual Normality	Jarque-Bera	stat=0.587, p=0.746 > 0.05	✓ PASS
Zero-Mean Residuals	Residual Mean	Mean = 0.000000	✓ PASS
Homoscedasticity	Breusch-Pagan	stat=15.02, p=0.182 > 0.05	✓ PASS
No Autocorrelation	Durbin-Watson	DW = 2.105 (range 1.5–2.5)	✓ PASS
No Multicollinearity	VIF (all predictors)	Max VIF = 2.478 < 5.0	✓ PASS
Overall Model Significance	F-test	F = 70.43, p < 0.001	✓ PASS
Cross-Validation Stability	10-Fold CV R^2	0.657 ± 0.107	✓ PASS

Table 15: Complete Regression Diagnostic Summary — All Assumptions Verified

All eight diagnostic criteria pass their respective tests, providing strong validation of the multivariate OLS regression framework applied to this dataset. The simultaneous satisfaction of normality, homoscedasticity, non-autocorrelation, and low VIF conditions confirms that the coefficient estimates are BLUE (Best Linear Unbiased Estimators) under the Gauss-Markov theorem, and that all associated t-tests, F-tests, and confidence intervals are statistically valid.

7. Discussion

7.1 The Dominant Role of Behavioral Engagement

The most consequential finding of this study is the exceptional explanatory power of behavioral engagement features: four LMS interaction variables alone account for 52.3% of variance in academic performance class — a remarkable result that establishes digital behavioral traces as the most information-rich predictor category in the model. This finding is consistent with decades of educational research on the attendance-achievement nexus (Tinto, 1987; Bean, 1985) but extends it to the domain of digital engagement metrics in LMS environments.

The near-equal standardized effect sizes of Visited Resources ($\beta^* = +0.202$) and Raised Hands ($\beta^* = +0.181$) suggest that both passive content engagement (viewing learning materials) and active social participation (raising hands, indicating questions) are independent and complementary pathways to academic success. This distinction has important pedagogical implications: instructors seeking to improve student outcomes should design interventions that simultaneously increase both resource accessibility and active classroom participation opportunities — not merely one or the other.

The Discussion variable, while positively correlated with performance (r = +0.308) and statistically significant in isolation, shows a comparatively small β* of +0.053 in the full model. This attenuation — from a moderate univariate correlation to a small multivariate coefficient — reflects suppression by the Visited Resources and Raised Hands variables, which absorb much of Discussion's explanatory variance. Students who are highly engaged in resources and classroom participation also tend to discuss more, making Discussion's independent contribution modest when controlling for the other behavioral dimensions.

7.2 Absence as the Most Powerful Single Determinant

The emergence of Student Absence as the single largest standardized coefficient (β* = −0.380) in the full model — exceeding even the strongest behavioral engagement predictors — is striking given that absence is added as the fourth block and therefore enters the model controlling for all behavioral, parental, and demographic predictors. This implies that absence has a direct, independent effect on academic performance beyond its correlation with low resource visits and low classroom participation. Students who are frequently absent miss instruction, fall behind on material, and lose the social engagement cues that drive platform interaction — a compounding disadvantage that the β* magnitude captures precisely.

The chi-square analysis reinforces this: among students absent more than 7 days, 91.3% fall in the Low performance class, and only 2.8% of High achievers are frequently absent. The practical implication is clear: absence is the most actionable single intervention target. School-based early warning systems designed to flag at-risk students should assign the highest priority weight to attendance monitoring — a recommendation directly supported by β* = −0.380 in the regression model.

7.3 Gender and Parental Effects

Female students outperform male students at the High achievement level (42.9% vs. 22.0%), yielding a Pearson correlation of r = +0.264 (female = positive) and a standardized coefficient of β* = +0.077 in the full model. This gender achievement gap — with female students performing higher — is consistent with the international literature on gender and academic achievement in secondary education, where female students have demonstrated consistently higher grade point averages than male peers in most subjects and countries since the 1980s.

Parent Answering Survey (β* = +0.137) retains the fourth-largest standardized effect in the full model, confirming that parental engagement with school communication systems has an independent positive effect on academic performance beyond the student's own behavioral engagement. This is an important finding for school policy: parent-school communication programs (surveys, meetings, progress reports) appear to generate academic benefits beyond their direct information provision function — possibly through motivational and accountability mechanisms operating within the family environment.

7.4 Limitations

The study carries important limitations. The ordinal encoding of the three-class outcome (L=1, M=2, H=3) assumes equal interval distances between classes, which may not hold precisely — though this is a standard and widely accepted practice for regression with ordered categorical outcomes. A multinomial logistic regression or ordinal regression (proportional odds model) would provide a more theoretically appropriate treatment of the three-class outcome, at the cost of interpretive simplicity.

The dataset's geographic scope — primarily Middle Eastern students in a single LMS platform — limits generalizability to different cultural, institutional, and technological contexts. The absence of socioeconomic status variables (parental income, educational attainment, household resources) is a notable omission: international educational research consistently identifies SES as a major determinant of academic achievement, and its absence from the model likely results in omitted variable bias for some of the parental engagement coefficients.

Finally, the cross-sectional design precludes causal inference: the observed associations between behavioral engagement and performance may reflect reverse causality (high-performing students engage more because they are succeeding), selection effects (inherently motivated students both engage more and perform better), or confounding by unmeasured personality and motivational traits.

8. Conclusion

This study has delivered a rigorous, diagnostically validated hierarchical multivariate regression analysis of academic performance determinants using the xAPI Learning Analytics Education dataset. The principal empirical findings are:

- The full multivariate OLS regression model achieves $R^2 = 0.6849$ and Adjusted $R^2 = 0.6660$ ($F = 70.43$, $p < 0.001$), explaining 68.5% of variance in academic performance — with all regression assumptions satisfied across normality (SW $p = 0.808$), homoscedasticity (BP $p = 0.182$), non-autocorrelation (DW = 2.105), and multicollinearity (max VIF = 2.478) diagnostics.
- Behavioral engagement alone accounts for 52.3% of variance (Model 1: $R^2 = 0.523$), establishing LMS interaction behaviors as the dominant academic performance predictor category — more powerful than parental, demographic, or absence factors independently.

- Student Absence is the single strongest standardized determinant ($\beta^* = -0.380$), adding 9.05 percentage points of R^2 independently of all other predictors. Among students absent more than 7 days, 91.3% fall in the Low class versus only 2.8% of High achievers.
- Visited Resources ($\beta^* = +0.202$) and Raised Hands ($\beta^* = +0.181$) are the dominant positive behavioral predictors; Parent Answering Survey ($\beta^* = +0.137$) is the strongest parental predictor; and female gender ($\beta^* = +0.077$) reflects a consistent performance advantage for female students.
- Parental involvement adds $\Delta R^2 = 0.048$ over behavioral engagement alone, confirming independent family-level effects on academic performance beyond students' own engagement behaviors.
- 10-fold cross-validation confirms model generalizability: CV $R^2 = 0.657 \pm 0.107$, with test RMSE = 0.374 — less than half a class unit of prediction error on average.

The findings have direct policy implications. Early warning systems should weight attendance most heavily, followed by resource visit frequency and classroom participation monitoring. Parent engagement programs — particularly structured survey and communication mechanisms — appear to generate independent academic benefits and should be maintained and expanded. Gender-differentiated interventions may be warranted to address the observed male underperformance gap. Future research should incorporate socioeconomic variables, employ longitudinal designs to establish causal direction, and extend the hierarchical regression framework to multinomial logistic or structural equation modeling formulations for more nuanced multi-class outcome analysis.

9. References

[1] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. International Journal of Database Theory and Application, 9(8), 119–136.

[2] Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review, 84(2), 191–215.

[3] Bean, J. P. (1985). Interaction effects based on class level in an explanatory model of college student dropout syndrome. American Educational Research Journal, 22(1), 35–64.

[4] Bronfenbrenner, U. (1979). The Ecology of Human Development. Harvard University Press.

[5] Chang, H. N., & Romero, M. (2008). Present, Engaged, and Accounted For: The Critical Importance of Addressing Chronic Absence in the Early Grades. National Center for Children in Poverty.

[6] Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). Equality of Educational Opportunity. Government Printing Office.

[7] Jeynes, W. H. (2007). The relationship between parental involvement and urban secondary school student academic achievement: A meta-analysis. Urban Education, 42(1), 82–110.

[8] Suleiman, M., Tukur, A., & Ibrahim, M. (2024). Key factors influencing students' academic performance: A systematic review. Journal of Engineering and Science in Information Technology, Springer JESIT.

[9] Tinto, V. (1987). Leaving College: Rethinking the Causes and Cures of Student Attrition. University of Chicago Press.

[10] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate Data Analysis (8th ed.). Cengage Learning.

[11] Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics (5th ed.). SAGE Publications.

[12] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.