Academic Performance Analysis: Interplay of Socioeconomic and Lifestyle Factors

Prathamesh Pradeep Khole, Nishant Kishor Khanorkar
University of California Santa Cruz

Abstract

Student academic achievement results from a complex interaction of contextual, behavioral, and socioeconomic factors. However, most research has been limited to predicting student outcomes using test scores and previous grades. Through the creation of informative statistical models, this research seeks to assess achievement in secondary school pupils in Portugal from multiple perspectives. It is possible to measure the ways in which various elements shape success when large-scale data on demographics, education, activities, attitudes, and technology use are integrated. The findings provide important behavioral and contextual information on learners, enabling parents, educators, politicians, and schools to make targeted reforms. Predictive indicators better show trends connecting student outcomes to whole academic experiences than historical scores alone. This analytical method creates fair solutions that are suited to fostering teenage education by delving deeper into the real-world interactions that shape it.

KEY WORDS: Exploratory Data Analysis, Predictive Modeling, Regression, Classification

1. Data Overview

This data [1] was collected in 2008 to study the high early school leaving rate in Portugal (40%) for 18 to 24-year-olds since it was higher than the European Union average value of just 15% (Eurostat 2007).

Our analysis will rely on a dataset [1] sourced from the students of Gabriel Pereira and Mousinho da Silveira schools; the data presents a mosaic of academic records enriched with demographic, social, and behavioral nuances for 649 students, represented by 30 unique features across three critical academic checkpoints — G1 (First-period), G2 (Second-period), and G3 (Final) grades for Portuguese and Mathematics.

Details of the key variables are as follows:

- **School Identifier (SI):** Binary variable representing the student's school, 'GP' for Gabriel Pereira and 'MS' for Mousinho da Silveira.
- **Demographic Profile (DP):** This includes binary variables such as sex ('F' for female, 'M' for male) and address type ('U' for urban, 'R' for rural), as well as numeric variables like age.
- **Family Background (FB):** Categorical and numeric variables such as family size ('LE3' for less or equal to 3, 'GT3' for more than 3), parent's cohabitation status, and parent's education level.
- **Educational Factors (EF):** Comprising nominal and numeric data, this set includes the mother's and father's jobs, the reason for choosing the school, and the student's travel time and weekly study time.
- **Personal Behaviors (PB):** Variables include participation in extracurricular activities, nursery education, desire for higher education, internet access at home, and romantic relationship status.
- **Health and Lifestyle (HL):** This includes assessments of the student's health status, alcohol consumption on weekdays and weekends, and school absences.
- **Academic Grades (AG):** Numeric variables indicate the student's first-period grade (G1), second-period grade (G2), and the final grade (G3), which is the target variable for many predictive models.

Our dataset does more than just track student grades; it provides a detailed look at how school activities and personal choices affect students' education. We will use this rich data to understand how students' backgrounds and decisions impact their academic outcomes, using statistical models to uncover the patterns within.

2. Prior Data Analysis

Cortez and Silva's study [2] on predicting student grades used data mining with Decision Trees, Random Forests, Neural Networks, and SVMs. They considered academic records, demographics, and social factors, confirming the strong influence of past performance and revealing the significance of absences and parental background. Suggesting real-time predictive systems in schools, they advocate for larger datasets and advanced feature selection to refine models. Kotsiantis et al. (2004) [3] corroborated the predictive power of academic history, highlighting the need for broader sociological investigations into these influences.

3. The Objective

Our primary goal is to investigate how socioeconomic factors and lifestyle choices, rather than previous grades

collectively impact academic outcomes in the Portuguese secondary educational context. We will analyze the intricate statistical relationships among the different elements of our dataset to gain a deeper understanding of this phenomenon.

Specifically, through our analysis we plan to:

1. Study the statistical relationships between students' grades and a series of independent variables, including socioeconomic factors, study habits, and personal lifestyle choices.
2. Evaluate the relative impact of various lifestyle choices and demographic factors on students' academic performances.
3. Predict the whether a student will pass or fail given certain socioeconomic factors and lifestyle choices they have, rather than relying on their previous grades G1 and G2.
4. Debunk popular hypotheses for school students.

Through rigorous analysis, we aim to provide both a qualitative and quantitative narrative on how external and personal factors influence educational achievements. The primary objective is to construct an in-depth analytical model that rigorously examines and evaluates hypotheses concerning the factors contributing to academic successes and failures. This model aims to elucidate how variations in specific determinants may have led to alternate educational results.

4. Exploratory Data Analysis (EDA)

Following is our approach to understand the influence of lifestyle factors on the educational outcome, discovering possible trends or relations between features.

4.1 Analyzed data trends and distributions:

We created histograms and bar plots for variables other than G1, G2, and G3 to understand their general trends and distributions.

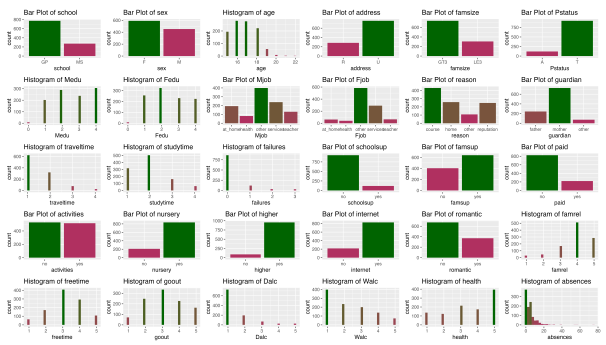


Figure 1: Histogram for all features except G1, G2, G3

Based on the histogram in Figure 1, we observe that many of our predictors are categorical, and some are skewed towards a particular category.

4.2 Identified data correlations:

We created multiple correlation heat maps to visualize potential relationships between different data points.

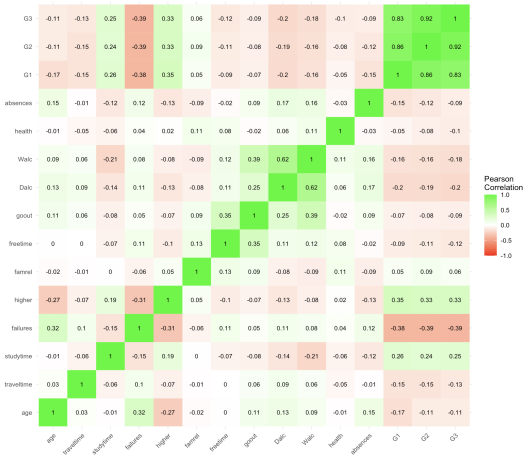


Figure 2: Feature correlation plot: All students

From above correlation maps in Figure 2, we can see that some of the predictors are highly correlated, and some have a strong influence on the final Grade (G3), and also on period Grade (G1 and G2), features failures, higher, Dalc, Walc, freetime, goout, traveltime, absents, and health seem to have strong influence on the final grades G3 (outcome).

4.3 Optimized model features:

We selected and removed highly correlated features to improve the final model's performance.

Figure 3: Heat map for highly correlated features

Next we identified these highly correlated features such as G1, G2, and G3, mothers and fathers education, and

aspiring for higher education with grades, this could allow us to remove some of these without losing their predictive power.

4.4 Explored final grade trends based on previous observations

Based on our observations, we analyzed and documented trends in final grades.

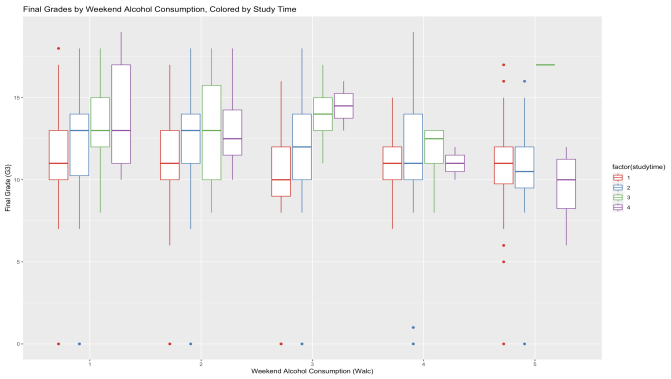


Figure 4: Final Grade ~Study time + Weekend Alcohol Consumption

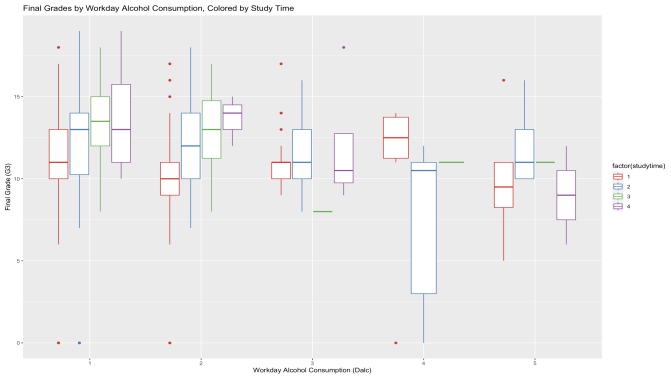


Figure 5: Final Grade ~Study time + Workday Alcohol Consumption

As workday alcohol consumption increases, there is a visible trend of lower median final grades, and this effect appears to be more pronounced for students with less study time.

The variability in final grades seems to increase with higher levels of weekend alcohol consumption, and similar to workday consumption, students with higher study times generally maintain higher median grades up to certain levels of alcohol consumption.

4.5 Investigated student lifestyle impacts:

We investigated the relationship between final grades and the lifestyle choices made by students.

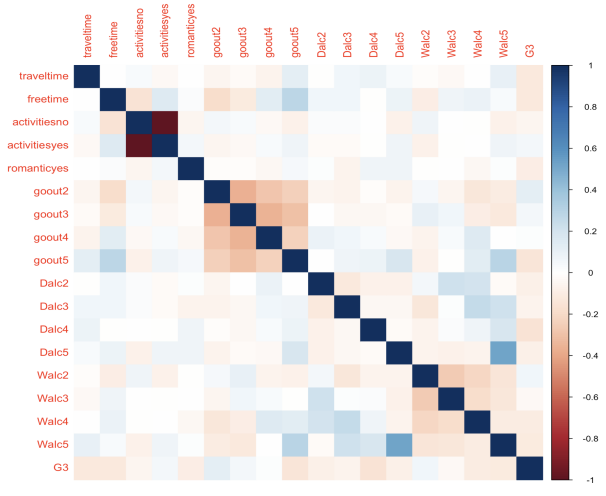


Figure 6: Correlation between lifestyle choices and final grade

4.6 EDA Summary and Observations:

- Elevated alcohol consumption weekday links to substantially lower grades, connecting drinking and achievement.
- Romantic relationship status correlates with lower performance, coupled students scoring below single peers.
- Learners with higher education aspirations earn improved grades over those lacking similar academic plans.
- Opting into supplementary paid lessons associates with higher scores versus compulsory coursework alone.
- Quantifying social interaction impacts on achievement requires further inquiry given model unpredictability.
- School absences negatively trend with performance but effect magnitude needs rigorous analysis.
- Parental contexts like occupations and engagement surface as potential student advancement drivers needing confirmation.

In summary, initial observations reveal understudied connections between adolescent behaviors, environments and academic milestones, preliminary mapping experiences to achievement.

5. Statistical Models

5.1 Multiple Linear Regression (MLR)

The Multiple Linear Regression (MLR) model predicts a continuous outcome variable ($G3$) as a linear combination of predictor variables. It is given by the following

equation:

$$G3 = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n + \epsilon$$

where $G3$ is the final grade, β_0 is the intercept, $\beta_1, \ldots, \beta_n$ are the coefficients for each predictor variable $X_1, \ldots, X_n$, and ϵ is the error term. The model assumes that the relationship between the dependent variable and the predictors is linear.

5.2 Logistic Regression

Logistic Regression is used when the outcome variable is binary. It estimates the probability that the outcome belongs to a particular category. For a binary outcome derived from the final grade ($G3$), the Logistic Regression model is expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_n \cdot X_n$$

where p is the probability of the outcome (e.g., passing the final grade), $\frac{p}{1-p}$ is the odds ratio, and $\beta_0, \beta_1, \ldots, \beta_n$ are the model coefficients corresponding to each predictor $X_1, \ldots, X_n$.

6. Fitted models for analysis

Four shortlisted models designed as below help understand how various socioeconomic factors and lifestyles choices influence a students final grade.

While some of the features are not significant and features like G1 and G2 are strong predictors of G3 owing to strong positive correlation, we are more interested in models that capture the influence of socioeconomic and lifestyle factors. So, we consider models without grade features G1 and G2 as they overshadow the minuscule influence of other features on the final grade which perfectly fits the scope of this research.

6.1 Model based on inferred features (M1)

Based on the Exploratory Data Analysis performed in figure 6, we observed that features traveltime, freetime, romantic interest, going out, daily and weekend alcohol consumption seem have high influence on the final grades and are those socioeconomic features which students choose, thus their lifestyle choices, to test this we created our first linear model (M1).

Model M1:
$G3 = \beta_0 + \beta_1 \times \text{traveltime} + \beta_2 \times \text{freetime} + \beta_3 \times \text{activities} + \beta_4 \times \text{romantic} + \beta_5 \times \text{goout} + \beta_6 \times \text{Dalc} + \beta_7 \times \text{Walc} + \varepsilon$

Based on the model M1 (table 1), we can observe that features free time, travel time, activities, romantic interest are significant in influencing the grades, where daily alcohol consumption is significant and it negatively influences the final grades, weekend alcohol consumption seems to not have a significant influence on the grades.

Table 1: Regression Coefficients for Model M1

Term	Estimate	Pr(>|t|)
(Intercept)	12.37203	< 2e-16 ***
traveltime	-0.38441	0.019953 *
freetime	-0.31776	0.011891 *
activitiesyes	0.54748	0.027160 *
romanticyes	-0.48214	0.058374 .
goout2	1.99044	0.000132 ***
goout3	1.62632	0.001313 **
goout4	1.75813	0.001117 **
goout5	1.10326	0.057043 .
Dalc2	-0.63478	0.078517 .
Dalc3	-0.52984	0.354550
Dalc4	-2.73486	0.000869 ***
Dalc5	-1.31748	0.163505
Walc2	-0.04144	0.901118
Walc3	-0.36033	0.344161
Walc4	-0.72726	0.118720
Walc5	-0.50079	0.471284

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

6.2 Model based on significant features for Mathematics grade (M2)

Next we decided to use all the variables and reduce the variables step-by-step using the stepAIC in both directions to obtain a model with actual significant features based on the AIC value for the model for predicting the final grade for Mathematics.

Model M2:
$G3 = \beta_0 + \beta_1 \times \text{sexM} + \beta_2 \times \text{age} + \beta_3 \times \text{famsizeLE3} + \beta_4 \times \text{Medu} + \beta_5 \times \text{Mjobhealth} + \beta_6 \times \text{Mjobother} + \beta_7 \times \text{Mjobservices} + \beta_8 \times \text{Mjobteacher} + \beta_9 \times \text{studytime} + \beta_{10} \times \text{failures} + \beta_{11} \times \text{schoolsupyes} + \beta_{12} \times \text{famsupyes} + \beta_{13} \times \text{romanticyes} + \beta_{14} \times \text{freetime} + \beta_{15} \times \text{goout} + \beta_{16} \times \text{absences} + \varepsilon$

The above model is the result of performing feature elemination in both directions starting with all the features other than G1 and G2 to predict G3 for Mathematics.

Based on the analysis of this model M2 (table 2), we can conclude that for final Mathematics grades of students features such as Gender, mothers education and study time, are significant and have a positive influence and no. of failures, romantic interests, goout, and absences are significant but have a negative influence.

Table 2: Regression Coefficients for Model M2

Term	Estimate	Pr(>|t|)
(Intercept)	13.67213	3.18e-05 ***
sexM	0.96171	0.03758 *
age	-0.28634	0.11395
famsizeLE3	0.72802	0.11754
Medu	0.55202	0.03382 *
Mjobhealth	1.47081	0.14707
Mjobother	-0.18623	0.77867
Mjobservices	0.97452	0.18572
Mjobteacher	-0.84531	0.38140
studytime	0.57107	0.03200 *
failures	-1.86045	1.96e-09 ***
schoolsupyes	-1.27767	0.05034 .
famsupyes	-0.82144	0.06597 .
romanticyes	-1.09244	0.01636 *
freetime	0.31303	0.16122
goout	-0.54499	0.00585 **
absences	0.05688	0.03587 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3: Regression Coefficients for Model M3

Term	Estimate	Pr(>|t|)
(Intercept)	8.90516	5.28e-07 ***
schoolMS	-1.51318	5.59e-10 ***
sexM	-0.57091	0.015726 *
age	0.16711	0.092231 .
Medu	0.30127	0.002454 **
guardianmother	-0.45308	0.073592 .
guardianother	0.03407	0.946911
studytime	0.40872	0.002580 **
failures	-1.48437	2.01e-13 ***
schoolsupyes	-1.33575	0.000196 ***
higheryes	1.86377	9.99e-07 ***
romanticyes	-0.42199	0.060679 .
Dalc	-0.35842	0.003584 **
health	-0.17961	0.014826 *
absences	-0.03687	0.126848

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

6.3 Model based on significant features for Portuguese grade (M3)

To compare if there are differences in the way lifestyle and socioeconomic factors influence final grades for different subjects, we again used all the variables and reduced the number of variables step-by-step using the stepAIC in both directions to obtain a model with actual significant features based on the AIC value for the model for predicting the final grade for Portuguese.

Model M3:
$G3 = \beta_0 + \beta_1 \times \text{schoolMS} + \beta_2 \times \text{sexM} + \beta_3 \times \text{age} + \beta_4 \times \text{Medu} + \beta_5 \times \text{guardianmother} + \beta_6 \times \text{guardianother} + \beta_7 \times \text{studytime} + \beta_8 \times \text{failures} + \beta_9 \times \text{schoolsupyes} + \beta_{10} \times \text{higheryes} + \beta_{11} \times \text{romanticyes} + \beta_{12} \times \text{Dalc} + \beta_{13} \times \text{health} + \beta_{14} \times \text{absences} + \varepsilon$

For Portuguese grades (table 3) we see that significant features have changed, while some have remained, other new ones have been added such as gurdian, schoolsup, health, and higher.

Based on this we observe that the lifestyle choices which can significantly influence grades for the two subjects are not same and some which are significant in for Mathematics are not significant for Portuguese and other way around.

6.4 Model based on significant features for Final grade (M4)

Finally we made a model by taking in all features and then eliminating them through stepAIC in both directions to get a model with least AIC and most significant features to predict the final grade G3, without taking into account period grades G1 and G2, and for all students regardless of their subject.

Model M4:
$G3 = \beta_0 + \beta_1 \times \text{schoolMS} + \beta_2 \times \text{addressU} + \beta_3 \times \text{famsizeLE3} + \beta_4 \times \text{Medu} + \beta_5 \times \text{Mjobhealth} + \beta_6 \times \text{Mjobother} + \beta_7 \times \text{Mjobservices} + \beta_8 \times \text{Mjobteacher} + \beta_9 \times \text{Fjobhealth} + \beta_{10} \times \text{Fjobother} + \beta_{11} \times \text{Fjobservices} + \beta_{12} \times \text{Fjobteacher} + \beta_{13} \times \text{studytime} + \beta_{14} \times \text{failures} + \beta_{15} \times \text{schoolsupyes} + \beta_{16} \times \text{paidyes} + \beta_{17} \times \text{higheryes} + \beta_{18} \times \text{internetyes} + \beta_{19} \times \text{romanticyes} + \beta_{20} \times \text{famrel} + \beta_{21} \times \text{goout} + \beta_{22} \times \text{health} + \varepsilon$

When taking into account just the final grades regardless of the subject, we get M4 (table 4). We see most of the significant features are common between the earlier two models (M2 and M3), while highlighting some new features such as paid classes.

7. Logistic Regression Model

To predict if a student will pass or fail based on socioeconomic factors and lifestyle choices, we utilize a logistic regression model. This method is ideal for binary outcomes like pass/fail, as it calculates the probability of passing from specific predictors, quantifying the impact of various factors on academic success.

We use the following model for our problem:

Model M5 (Logistic Regression):
$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \text{schoolMS} + \beta_2 \times \text{Fedu} + \beta_3 \times \text{Mjobhealth} + \beta_4 \times \text{Mjobother} + \beta_5 \times \text{Mjobservices} + \beta_6 \times \text{Mjobteacher} + \beta_7 \times \text{traveltime} + \beta_8 \times \text{failures} + \beta_9 \times \text{schoolsupyes} + \beta_{10} \times \text{paidyes} + \beta_{11} \times \text{activitiesyes} + \beta_{12} \times \text{higheryes} + \beta_{13} \times \text{health} + \beta_{14} \times \text{absences} + \beta_{15} \times \text{Walc}$

Table 4: Regression Coefficients for Model M4

Term	Estimate	Pr(>|t|)
(Intercept)	9.86455	< 2e-16 ***
schoolMS	-0.40530	0.137932
addressU	0.39536	0.125448
famsizeLE3	0.38296	0.103361
Medu	0.19787	0.124438
Mjobhealth	1.17489	0.026538 *
Mjobother	0.01799	0.954266
Mjobservices	0.61785	0.100463
Mjobteacher	-0.04271	0.931004
Fjobhealth	-0.05961	0.933532
Fjobother	-0.15477	0.739696
Fjobservices	-0.49365	0.309206
Fjobteacher	0.91146	0.152413
studytime	0.44517	0.000826 ***
failures	-1.75839	< 2e-16 ***
schoolsupyes	-1.34244	9.47e-05 ***
paidyes	-1.05565	7.76e-05 ***
higheryes	1.45195	0.000417 ***
internetyes	0.41283	0.145155
romanticyes	-0.60777	0.007129 **
famrel	0.19332	0.093890 .
goout	-0.27957	0.002776 **
health	-0.21207	0.005452 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The model is obtained by taking into account all the features (except G1 and G2) and then performing feature selection by using stepAIC to remove the less significant features in each iteration and also reduce the model AIC at each step to obtain the model with the least possible AIC.

The model is used to predict the final grade using a newly defined column 'Pass', which is set to 1 (Yes) else 0 (No) for every student having G3 >= 12 (Passing grade).

Model M5 (table 5) also gives significance to features similar to the earlier models (M2, M3 and M4) with addition of a few new features such as fathers education and Weekend alcohol consumption.

8. Model evaluation and selecting the best model

Confusion Matrix:

		Reference		
		Fail	Pass	Total
Predicted	Fail	89	45	134
	Pass	16	58	74
	Total	105	103	208

Key Statistics:

Table 5: Logistic Regression Coefficients for Model M5

Term	Estimate	Pr(>|z|)
(Intercept)	-0.84963	0.142844
schoolMS	-0.68919	0.000534 ***
Fedu	0.22565	0.005356 **
Mjobhealth	0.98047	0.006517 **
Mjobother	0.71015	0.002954 **
Mjobservices	0.84371	0.001747 **
Mjobteacher	0.33169	0.283013
traveltime	-0.22345	0.061928 .
failures	-1.29192	4.32e-09 ***
schoolsupyes	-1.38122	2.47e-07 ***
paidyes	-0.68349	0.000448 ***
activitiesyes	0.31540	0.051403 .
higheryes	1.58731	0.000251 ***
health	-0.09478	0.101750
absences	-0.05488	0.000431 ***
Walc	-0.15970	0.014856 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Accuracy:	0.7067
Correct Predictions:	147
Test Size:	208

Area under the curve: 0.7816

Likelihood Ratio Test for Logistic Model Models:

- Model 1: Pass 1
- Model 2: Pass (school + sex + age + address + famsize + Pstatus + Medu + Fedu + Mjob + Fjob + reason+guardian+traveltime+studytime+failures+ schoolsup + famsup + paid + activities + nursery + higher + internet + romantic + famrel + freetime + goout + Dalc + Walc + health + absences + G1 + G2 + G3) − G1 − G2 − G3

Table 6: Deviance Table:

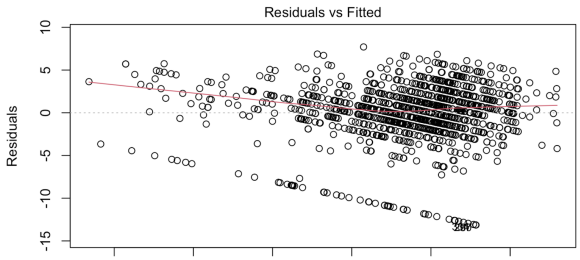
	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
M1	835	1158.36			
M2	796	902.94	39	255.42	< 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

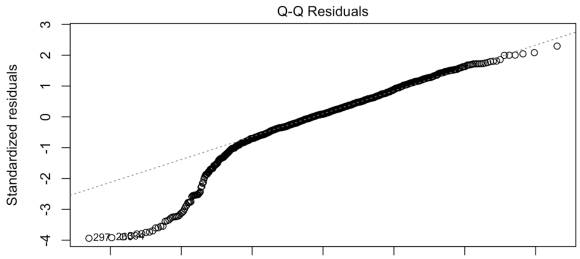
Based on the Likelihood ratio test where model M2 (with parameters) is significant compared M1 (without parameters), and the accuracy and area under the curve we can conclude that our model has fit the data well.

9. Residual Analysis

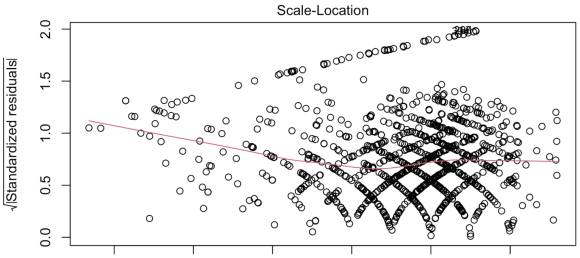
Below plots demonstrate residual analysis for Multiple Linear Regression model explained earlier:



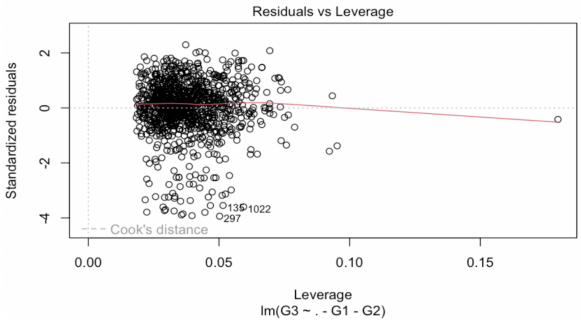
The Residuals vs Fitted Plot shows some patterns that might suggest non-linearity in the relationship between predictors and the response variable, or it may indicate the presence of heteroscedasticity. The **non-random** pattern in the residuals is due to the nature of response variable being a grade in between 0 to 20. As seen from the earlier plots, the relationship between the response and predictors is not linear, causing a pattern in the residuals when imposing a linear model.



The Normal Q-Q Plot indicates that residuals mostly follow a normal distribution but with some deviations at the tails. This could suggest that the residuals are not perfectly normal, which is common in real-world data but can affect inference.



The Scale-Location Plot shows some signs of non-constant variance (heteroscedasticity), as the residuals don't seem to be randomly scattered around the horizontal line.



The Residuals vs Leverage Plot doesn't indicate any points with high leverage that are also outliers in the response (which would be outside Cook's distance lines).

10. Results and Conclusion

Analysis Reveals Multi-Faceted Drivers Impacting Academic Performance. Several insightful conclusions emerge from our in-depth analytical modeling regarding the factors that collectively shape academic achievement among adolescent students:

- First, the data indicates a significant negative relationship between alcohol consumption levels and final grades. Escalated drinking levels correlate with poorer performance in school.
- Additionally, our findings conclude romantic involvement during secondary schooling also bears a detrimental link to educational outcomes. Students invested in relationships scored disproportionately lower than their single peers.
- Meanwhile, nurturing ambitious academic mindsets centered on aspirations for higher education bore positive dividends. Learners setting advanced scholarly goals excelled relative to those lacking postsecondary educational plans.
- Pursuing extra, paid supplementary coursework also contributed to boosted academic results. Students undertaking additional lessons demonstrated heightened performance versus those relying solely on compulsory studies.
- Delving deeper, certain lifestyle factors proved more nebulous in their connections to achievement. Models linking social activities to grades yielded more inconsistent patterns, necessitating further data to discern effects.
- Similarly, while increased school absences associate directionally with poorer rankings, continuing research must solidify the substantive sway of missed school on performance.
- Finally, familial environments prominently assist in cultivating student success. Maternal educational backgrounds and vocations in healthcare link to enhanced outcomes. The models quantify these maternal investments in development as strongly beneficial.

In summary, an intricate, interwoven array of decisions students make with their time, goals, relationships and home support substantially impact their academic trajectories. Our analyses illuminate this rarely quantified connection between lifestyle choices and success.

References

[1] Paulo Cortez. Student Performance. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5TG7T.

[2] P. Cortez and A. M. Gonçalves Silva. Using data mining to predict secondary school student performance. 2008.

[3] Sotiris Kotsiantis, Christos Pierrakeas, and P. Pintelas. Predicting students'performance in distance learning using machine learning techniques. *Applied Artificial Intelligence*, 18:411–426, 01 2004.

[4] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[5] Jim Albert and Maria Rizzo. *Introduction*, pages 1–42. Springer New York, New York, NY, 2012.