Modeling Academic Performance: A Multiple Linear Regression Analysis of High School Students' GPA

Kristine Ann L. Bihag Kim Therese M. Bosoboso Jerahmeel T. Lim
Marie Rose Telly C. Lopez Benedicto M. Merales Jr.

ABSTRACT. Academic performance is a critical measure of student performance and a key indicator of educational system effectiveness. This study employs multiple linear regression (MLR) to explore the influence of demographic characteristics, study habits, parental involvement, and participation in extracurricular activities on the Grade Point Averages (GPA) of two thousand, three hundred ninety-two (2,392) high school students aged 15-18. Key predictors analyzed include age, gender, weekly study hours, parental support levels, and involvement in sports, music, and volunteering activities. The final regression model retained eight significant variables: absences, age, gender, music participation, sports participation, parental support, log-transformed study time, and tutoring. Results indicate that regular attendance, higher levels of parental support, and participation in music and sports positively influence GPA, while higher absences and lower parental support negatively impact academic performance. These findings offer meaningful guidance for educators and school administrators in crafting strategies to support student achievement and elevate overall educational performance. Additionally, these findings provide a foundation for future researchers to build upon, enabling them to explore more relationships and additional factors influencing academic performance.

I. INTRODUCTION

Students' academic performance represents a critical component within the constellation of factors that determine overall student success. It is not merely a measure of individual achievement but also a reflection of the effectiveness of the educational system. It is important to recognize that a person's life trajectory is often shaped by the knowledge they acquire and how effectively they apply it to benefit themselves, their community, and the world. This underscores the fundamental importance of education. Education plays a vital role in developing skilled human capital, driving economic growth, and addressing societal challenges (Tadese et al., 2022).

Ali et al. (2013) noted that students' previous educational outcomes serve as the most significant predictor of their future academic success. This implies that higher past performance is strongly associated with better academic achievements in future endeavors. Within this framework, academic performance emerges as a key indicator of both individual potential and institutional effectiveness.

High school, as a pivotal stage in the educational process, is characterized by complex interactions between a student's environment, habits, and support systems. This stage is particularly significant for researchers and educators aiming to understand and enhance student success, as it is where multiple factors converge to influence outcomes. A variety of factors—including study habits, parental involvement, demographics, and extracurricular activities, significantly impact the academic performance of high school students. A comprehensive understanding of these factors is essential for educators and researchers to improve student outcomes and inform effective teaching practices.

Demographic factors such as age, gender, and socio-economic status provide a critical context for academic performance, influencing students' engagement and capacity to excel in their studies. Similarly, effective study habits are widely recognized as a fundamental factor in shaping students' academic performance. Key indicators of study habits include the amount of time students dedicate to studying each week (Study Time Weekly), the number of absences they accumulate throughout the academic year (Absences), and whether they receive additional academic support through tutoring. These factors collectively provide valuable insights into students' learning behaviors and their potential impact on academic outcomes. In the study of Rabia et al. (2017), results showed that there is a significant relationship between study habits and the academic performance of students.

Parental involvement, characterized by the level of support and engagement in a child's education, has been shown to impact motivation, discipline, and overall academic performance. Additionally, extracurricular activities, such as participation in sports, music, and volunteering, contribute to the holistic development of students, fostering skills like time management, teamwork, and discipline, which often translate to better academic performance.

This study aims to explore how these interconnected factors–demographic characteristics, study habits, parental involvement, and extracurricular activities—jointly influence students' academic performance. By analyzing these variables, the research seeks to provide valuable insights into strategies for improving student success and shaping effective education methods. To investigate the relationships between these variables and academic performance, this study employs multiple linear regression analysis. This statistical technique allows for the simultaneous examination of the impact of multiple predictors on academic performance, providing a comprehensive understanding of how these interconnected factors collectively influence student performance. By leveraging this approach, the research seeks to uncover actionable insights that can inform strategies for enhancing educational practices and improving student success.

II. RELATED LITERATURE

Student performance is the evaluation of a student's participation, literacy, performance, and professionalism, either through formative or summative assessment (Dugan et. al, 2011). Measuring students' performance is critical for the advancement of education. By analyzing their progress, educational sectors can develop their techniques, materials, and other practices to be as efficient as possible. As a result, students are prepared to live a self-sufficient life as well as deal with political and social issues they might face (Kuh et. al, 2006). Kharoua (n.d.) provided a dataset where measured performance falls under the quantified attributes, also known as summative assessment. Summative assessment is the evaluation of performance based on results from quizzes, exams, reports, etc., and focuses on ranking and certification (Shafiq and Siddiquah, 2011). On the other hand, formative assessment is the evaluation based on people other than the students like the parents and teacher (Mansell and James, 2009). Shafiq and Siddiquah (2011) also suggest that summative assessment is an assessment of learning while formative assessment is an assessment. Thus, formative assessment should be used for enhancing student materials yet it is hard to quantify, so summative assessment is used but not as accurate (OECD Policy Beliefs, 2005). According to Yang and Li (2008), there are three ways that student success can be measured summatively: quantifying performance and non-performance related attributes by Student Attribute Matrix (SAM), classifying based on Back Propagation Neural Network (BP-NN) according to students' prior knowledge, and the causal relationship between the two. While society has adapted to the use of intelligence quotient as a main basis for an individual's cognitive competence, it might be unreliable since it can change over the development stages, especially adolescence (Sternberg, 2001). That's why constant evaluation through written summative assessment is crucial for accurate quantification of student performance (Deary et. al, 2007). In Kharoua's dataset, the dependent variable is GPA. According to Kuncel et. al

(2007), GPA or grade point average has a good predictive ability as tested on business school students, especially when combined with GMAT or Graduate Management Admission Test.

Although quantified student results are effective in reflecting actual academic performance (Chemers, et. al, 2001), Hardy (2014) suggests that schools have no full autonomy in relation to the academic results of students. This would pose a potential risk to the integrity of such numbers. This is because there are several factors that could impact performances positively or negatively that would not reflect on paper.

Richardson et. al (2012) and Cassady and Johnson (2002) provide evidence that emotions and high levels of anxiety would interfere with students' cognitive ability which may lead to poorer thought organization. Richardson conducted a meta-analysis of 13 years of research, finding that there are 5 emotional factors that affect performance: (a) personality traits, (b) motivational factors, (c) self-regulatory learning strategies, (d) students' approaches to learning, and (e) psychosocial contextual influences. Cassady and Johnson found out that cognitive test anxiety has a stable and negative impact on test scores by evaluating exams and self-reported performance on SATs for 168 undergraduate students.

Parental involvement may have various impacts depending on the context. Based on the analysis of the Area of Academic Achievements' study feature, more involvement suggests better performance when a global indicator of achievement like GPA is compared. On the other hand, based on the Parental Involvement Dimension, subject-specific parameters suggest a low relationship between involvement and performance (Fan and Chen, 2001).

It is crucial as well to take into account the potential benefit an older student may have over a young one (Voyles, 2011). Age differences among high school students in the same classroom have been a topic of significant interest in educational research, particularly regarding their impact on academic performance. A meta-analysis by La Paro and Pianta (2000) revealed that older children in school classrooms tend to outperform their younger, yet still age-appropriate peers academically. However, it is worth noting that while some researchers acknowledged the "short-term academic and behavioral benefits" of delayed school entry, there remains disagreement on its long-term advantages.

Another variable examined in this study is gender, which refers to the distinction between males and females and how this characteristic influences their attitudes, perceptions, and engagement with academic activities (Okoh, 2007). According to Buadi (2000), there is still no clear evidence linking gender to students' intellectual achievement. This ambiguity emphasizes how crucial it is to look into whether there are notable distinctions in the academic performance of male and female high school students. From the study of Çelikkol (2023), it cited Ramey et al. (1998) that girls tend to exhibit higher academic achievement and better adjustment levels compared to boys. Similarly, Cimen (2000) identified gender as a significant factor in psychosocial development, noting that female students are more likely to take on responsibilities and engage in independent activities than their male counterparts. Supporting this, Bahali et al. (2009) found that girls demonstrate more favorable outcomes in school maturity during the preschool period compared to boys. Furthermore, Kaya and Akgun (2016) observed that girls showed greater school adjustment than boys. These findings collectively highlight the consistent trend of girls demonstrating higher levels of academic performance, adjustment, and maturity compared to boys, emphasizing the importance of considering gender, along with the age of students, as key factors influencing academic performance.

Students' academic success is also correlated with the socioeconomic status of their parents, which includes their income, professional and academic credentials, and occupational connections (Ali et al., 2013). Students from higher socio-economic backgrounds tend to outperform those from lower socio-economic backgrounds. More stable finance leads to an evident rise in performance, since students have more accessibility to better materials and references (Reardon, 2018; Sirin 2005). According to Jeynes (2002), a student's socio-economic status is usually influenced by the income, occupation, and educational attainments of their parents. Given its significant influence, it is unsurprising that socio-economic status is one of the primary factors examined in research on academic performance prediction.

Similarly, the relationship between attendance and academic performance aligns with findings suggesting that students who attend lectures more frequently tend to achieve higher grades due to their intrinsic motivation and commitment to learning (Andrietti & Velasco, 2015). This correlation highlights the impact of absences on performance, as frequent absences may reflect lower motivation, affecting a student’s ability to grasp key concepts and excel in their studies. Moreover, the findings underscore that students with lower inherent ability or academic returns may require additional study effort to compensate for missed instructional time, further emphasizing the role of consistent attendance and effective study habits in achieving academic success.

Researchers use regression models to improve education by investigating certain phenomena like learning capability, attitude development, and fail-pass rates (Van Dusen and Nissen, 2019). Van Dusen and Nissen used disaggregation and hierarchical linear modeling to analyze hierarchical datasets as well as aggregation to collapse one set to another. It is concluded in the study that MLR is efficient in isolating variables of interest while considering other variables as well. Ding (2006) proves this claim as well by focusing on conventional regression analysis and regression mixture analysis in the field of education. Conventional regression analysis allows a set of estimated regression parameters to capture population characteristics in the sample. Regression mixture analysis, on the other hand, is a more flexible approach since it can identify regression data in varying latent classes. Despite the presence of statistical significance and signs of predictors, Niu (2017) claimed that logistic regression lacks certainty in interpreting relationship magnitude. The odds ratio is incorrectly interpreted as relative risk, misleading some conclusions. Despite this, Yang et. al claimed that using multiple linear regression (MLR) combined with principal component analysis (PCA) to establish a prediction model would be optimal for determining predictive performance. Yang et. al utilized predictive MSE (pMSE) and predictive mean absolute percentage correction (pMAPC) to balance the drawbacks of basic MLR which are coefficient of determination, mean square error, and the quantile-quantile plot since it cannot assess the MLR’s accuracy. According to Yang et. al’s findings, MLR is best for building a prediction model for student academic performance for the blended calculus course with many variables.

To summarize, student performance can be measured either using summative or formative assessment, where the former is used in building models since quantified values are easier to analyze (Yang et. al, 2018). Various factors can affect student performance such as: emotions, parental involvement, age, gender, socioeconomic status, parental education, parental involvement, study habits, and extracurricular activities. Using MLR with different tests is best for building a prediction model for student performance. While students’ summative performance can be predicted using regression analysis, formative performance is still best for improving the quality of education (OECD Policy Beliefs, 2005).

III. METHODOLOGICAL SKETCH

The data utilized in this research were sourced from an online platform called Kaggle. The Students Performance dataset from 2024 offers detailed insights into 2,392 high school students, covering their demographics, study patterns, parental engagement, participation in extracurricular activities, and academic outcomes. The dependent variable, ***Grade Point Average (GPA)***, categorizes students' grades into specific groups, making it a valuable resource for educational studies, predictive analytics, and statistical exploration.

a. DEFINITION OF VARIABLES

Dependent Variable (to edit

Grade Point Average (GPA) - This variable measures the students' academic performance and is assessed on a scale of 2.0 to 4.0, with higher values indicating greater academic achievement. GPA is a cumulative indicator of a student's grades across all disciplines that is influenced by a variety of factors including study habits, parental involvement, and participation in extracurricular activities. In this study, GPA will be analyzed in relation to other variables to identify potential correlations and underlying influences on academic outcomes.

Independent Variables

Demographic Details:

- ***Age (Age)*** - The students involved in this study ranged in age from 15 to 18 years. This range comprises the normal age group of high school students, making it an appropriate sample for studying academic performance and related characteristics throughout the important developmental period. Furthermore, the selection of this range shows the goal to investigate trends, behaviors, or outcomes that are common among students nearing the end of their secondary school.

- ***Gender (Gender)*** - This classifies the students according to their gender, with 0 being Female and 1 representing Male. This binary classification allows for the analysis of gender-specific trends, patterns, or gaps in academic achievement. The inclusion of this variable allows the study to investigate potential gender differences or correlations.
- ***Ethnicity (Ethnicity)*** - The variable Ethnicity categorizes the students with regards to their self-identified ethnic background, coded as:
    - 0: Caucasian
    - 1: African American
    - 2: Asian
    - 3: Other

It reflects the cultural and demographic diversity of the student population, which can be useful in assessing inequalities in educational outcomes, experiences, and access to resources. For regression analysis, this categorical variable was converted into dummy variables using *African American* as the reference category. Each of the remaining categories (Asian, Caucasian, and Other) was represented as a binary variable (1 for membership in the category and 0 otherwise).

- ***Parental Education (ParentalEducation)*** - This variable represents the educational level of the student's parents, coded as:
    - 0: None
    - 1: High School
    - 2: Undergraduate
    - 3: Bachelor’s Degree
    - 4: Higher Education

Parental education is often considered a substitute for socioeconomic status and can provide valuable insights into how a family background affects academic outcomes. For analysis purposes, this categorical variable was transformed into dummy variables as well to assess the distinct effects of each educational level on the dependent variable. In the case of this variable, *Bachelor’s* is the reference category.

Study Habits:

- ***Weekly Study Time (StudyTime)*** - This variable corresponds to the number of hours students dedicate to studying each week, ranging from 0 to 20 hours. By using a continuous scale, the study can capture variations in study habits and evaluate whether incremental increases in study time are associated with improved academic performance or outcomes.
- ***Absences (Absences)*** - This variable represents the total number of days a student has been absent from school during the academic year, with values ranging from 0 to 30. High levels of absences may indicate underlying issues such as health problems, lack of engagement, or personal challenges, whereas low or no absences can suggest strong attendance habits and consistent access to education.
- ***Tutoring (Tutor)*** - This variable demonstrates whether a student has received formal tutoring support, coded as dummy variables wherein 0 is for No Tutoring and 1 for Tutoring. Tutoring may include one-on-one or group sessions aimed at improving academic performance, addressing specific learning challenges, or providing enrichment opportunities.

Parental Involvement:

- ***Parental Support (ParentalSupport)*** - Parental support refers to the involvement, encouragement, and resources parents provide to their children across various aspects of life, including academics. Here, we categorize it into five levels:
  - 0: None
  - 1: Low
  - 2: Moderate
  - 3: High
  - 4: Very High

To analyze its impact, parental support was encoded using dummy variables, with *High* serving as the reference category. By coding parental support, educators and researchers can assess its impact and determine if students benefit from this kind of additional resources or interventions.

Extracurricular Activities:

- ***Extracurricular (Extracurricular)*** - Extracurricular participation, where 0 indicates No and 1 indicates Yes, encompasses all organized activities outside the standard curriculum, such as clubs, arts, and competitions.
- ***Sports (Sports)*** - Sports participation, where 0 is No and 1 is Yes, focuses on the athletic involvement of the students, promoting their physical health, discipline, and a sense of camaraderie.
- ***Music (Music)*** - Music participation, 0 is No and 1 is Yes, involves engaging in musical pursuits like playing instruments, singing, or dancing, which can foster creativity, improve cognitive skills, and provide emotional expression.
- ***Volunteering (Volunteering)*** - Volunteering participation, 0 for No and 1 for Yes, highlights engagement in community service or charitable activities, cultivating social responsibility and a sense of purpose.

b. MODEL SPECIFICATION

The proposed model is a multiple linear regression with GPA as the dependent variable and 12 indicators as independent variables.

$$GPA = \beta_0 + \beta_1 Absences + \beta_2 Age + \beta_3 Ethnicity + \beta_4 Extracurricular +$$

$$\beta_5 Gender + \beta_6 Music + \beta_7 ParentalEducation + \beta_8 ParentalSupport$$

$$+ \beta_9 Sports + \beta_{10} StudyTime + \beta_{11} Tutor + \beta_{12} Volunteering + \varepsilon$$

c. MODEL ASSUMPTIONS

- All tests and variable selection processes have a 0.05 level of significance.
- The GPA variable is treated as continuous and unbounded, recognizing its constrained range (2.0 to 4.0) as a potential limitation.
- The relationship between GPA (dependent variable) and the independent variables is assumed to be linear and subject to random error.

- The error terms are assumed to follow a normal distribution, exhibit homoscedasticity, and lack correlation unless evidence indicates otherwise.
- The independent variables are presumed to be linearly independent or do not demonstrate multicollinearity unless evidence indicates otherwise.
- There is no autocorrelation between the observations unless evidence indicates otherwise.

IV. RESULTS AND DISCUSSIONS

Initial Results in Multiple Linear Regression

All 12 predictor variables were regressed on the **GPA** as the dependent variable. The initial model exhibits strong explanatory power, with a high R^2 value of 0.9542, indicating that approximately 95.42% of the variance in GPA is accounted for by the predictors included in the model. Table 1 presents the parameter estimates for the initial model.

Table 1: Parameter Estimates of the Initial Model

	Estimate	Std. Error	t-value	p-value
(Intercept)	3.0692000	0.0628574	48.828	<2e-16 ***
Absences	-0.0997802	0.0004766	-209.350	< 2e-16 ***
Age	-0.0058189	0.0036017	-1.616	0.1063
EthnicityAsian	-0.0002023	0.0127221	-0.016	0.9873
EthnicityCaucasian	-0.0085257	0.0105504	-0.808	0.4191
EthnicityOther	-0.0065756	0.0159560	-0.412	0.6803
ExtracurricularYes	0.1907703	0.0083009	22.982	<2e-16 ***
GenderMale	-0.0151641	0.0080675	-1.880	0.0603 .
MusicYes	0.1421547	0.0101638	13.986	<2e-16 ***
ParentalEducationHigh School	0.0034774	0.0126238	0.275	0.7830
ParentalEducationHigher	0.0339480	0.0207757	1.634	0.1024

ParentalEducationNone	0.0164296	0.0163111	1.007	0.3139
ParentalEducationSome College	0.0180399	0.0121689	1.482	0.1384
ParentalSupportLow	-0.2960930	0.0116404	-25.437	<2e-16 ***
ParentalSupportModerate	-0.1530929	0.0104106	-14.706	<2e-16 ***
ParentalSupportNone	-0.4579696	0.0154679	-29.608	< 2e-16 ***
ParentalSupportVery High	0.1563675	0.0144689	10.807	<2e-16 ***
SportsYes	0.1934421	0.0087805	22.031	<2e-16 ***
StudyTime	0.0289142	0.0007141	40.489	<2e-16 ***
TutorYes	0.2500373	0.0087927	28.437	<2e-16 ***
VolunteeringYes	-0.0077734	0.0110883	-0.701	0.4833
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1				
Residual standard error: 0.1967 on 2371 degrees of freedom				
Multiple R-squared: 0.9542, Adjusted R-squared: 0.9538				
F-statistic: 2470 on 20 and 2371 DF, p-value: < 2.2e-16				

This shows that the overall initial model is highly significant, as evidenced by the F-statistic with p-value $< 2.2e - 16$, suggesting that the predictors collectively contribute meaningfully to explaining GPA. Key predictors such as Absences, Extracurricular participation, Music involvement, Parental support, Study time, and Tutoring show statistically significant relationships with GPA. Conversely, certain variables, such as age, volunteering, and parental education levels, exhibit non-significant effects since they have p-values greater than the conventional thresholds, suggesting limited contributions to explaining GPA variability in this dataset. These predictors may be considered for exclusion in subsequent model refinements

Diagnostic Checking and Validation of Assumptions

a. Linearity

The linearity assumption of the regression function was evaluated using residual plots. The plots of residuals against each predictor variable revealed no discernible patterns, indicating no significant departures from linearity.

b. Normality

The normality assumption of the error terms was assessed using the Shapiro-Wilk test, which produced a p-value of 0.01165, indicating a violation of normality at the 0.05 significance level since $0.01165 < 0.05$. Table 2 exhibits the results for the normality test.

Table 2: Normality Test for Initial Model

W	p-value	α
0.99826	0.01165	0.05

H_0: Data is normal vs. H_a: Data is not normal

To address this, a log transformation was applied to the predictor variable *StudyTime*, and a revised model was developed. The Shapiro-Wilk Test was subsequently performed on the new model to reassess the normality of the error terms. Figures 1 and 2 below show the residual plots of the transformed variable.

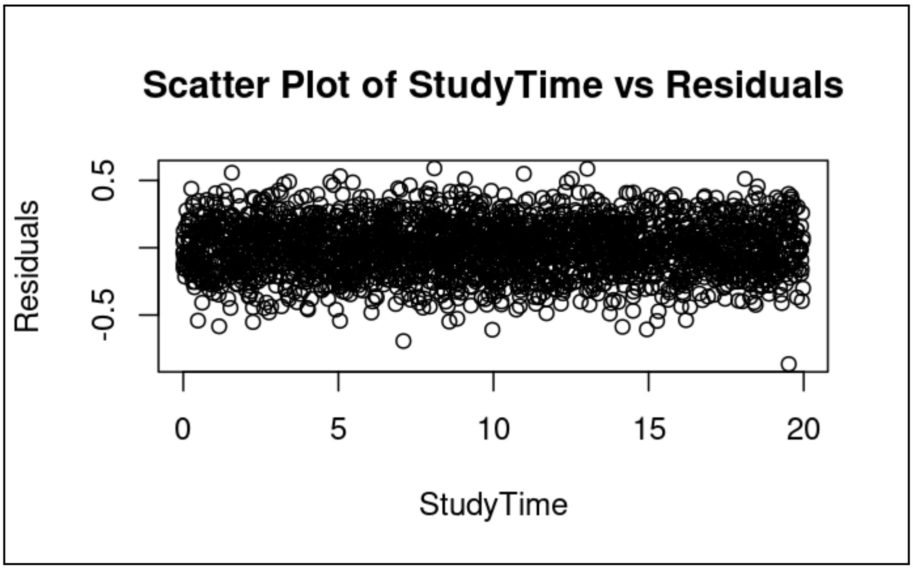

Figure 1: Scatter Plot of StudyTime vs First Model Residuals

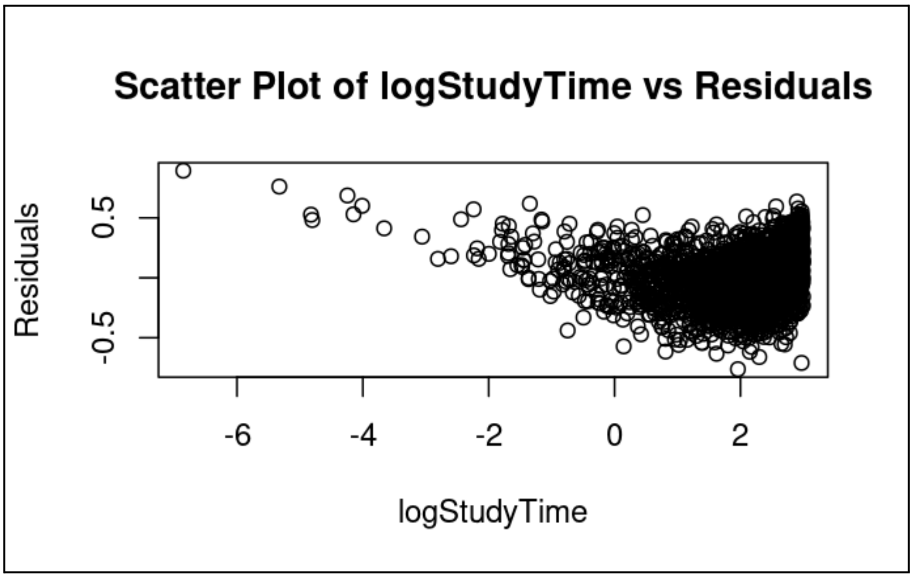

Figure 2: Scatter Plot of logStudyTime vs Second Model Residuals

Table 3: Normality Test for Second Model

W	**p-value**	α
0.99947	0.7762	0.05

H_0: Data is normal vs. H_a: Data is not normal

Based on the normality test results shown in Table 3, the p-value of 0.7762 exceeds the significance level of 0.05, suggesting that the current model satisfies the assumption of normality for the error terms. Despite achieving normality, variable selection was performed to identify significant predictor variables. The variable selection process produced consistent results, with Table 4 summarizing the outcomes of the backward elimination procedure.

Table 4: Summary of the Backward Elimination

Step	Variable Removed	Number of Variables In	R^2
1	Ethnicity	11	0.94583
2	ParentalEducation	10	0.94573

Backward elimination at a significance level of 0.05 determined 10 significant predictor variables. With this, the third model is given by

$$GPA = \beta_0 + \beta_1 Absences + \beta_2 Age + \beta_3 Extracurricular + \beta_4 Gender$$
$$+ \beta_5 Music + \beta_6 ParentalSupportLow + \beta_7 ParentalSupportModerate +$$
$$\beta_8 ParentalSupportNone + \beta_9 ParentalSupportVeryHigh + \beta_{10} Sports$$
$$Yes + \beta_{11} logStudyTime + \beta_{12} Tutor + \beta_{13} Volunteering$$

For the third model, multiple linear regression analysis was performed following the elimination process, with the results illustrated in Table 5.

Table 5: Parameter Estimates of the Third Model

	Estimate	Std. Error	t-value	p-value
(Intercept)	3.1017498	0.0670177	46.283	< 2e-16 ***
Absences	-0.0994338	0.0005172	-192.253	< 2e-16 ***
Age	-0.0068340	0.0039084	-1.749	0.0805 .
ExtracurricularYes	0.1912289	0.0090065	21.232	< 2e-16 ***
GenderMale	-0.0209629	0.0087637	-2.392	0.0168 *
MusicYes	0.1440673	0.0110159	13.078	< 2e-16 ***
ParentalSupportLow	-0.2981676	0.0126306	-23.607	< 2e-16 ***
ParentalSupportModerate	-0.1562982	0.0113049	-13.826	< 2e-16 ***
ParentalSupportNone	-0.4608616	0.0167907	-27.448	< 2e-16 ***
ParentalSupportVery High	0.1607270	0.0156908	10.243	< 2e-16 ***
SportsYes	0.1952474	0.0095318	20.484	< 2e-16 ***
logStudyTime	0.1379006	0.0043033	32.045	< 2e-16 ***
TutorYes	0.2503758	0.0095514	26.213	< 2e-16 ***
VolunteeringYes	-0.0132940	0.0120427	-1.104	0.2697
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 0.2138 on 2378 degrees of freedom				

```
Multiple R-squared:  0.9457,   Adjusted R-squared:  0.9454
F-statistic:  3188 on 13 and 2378 DF,  p-value: < 2.2e-16
```

Upon deriving the third model, the normality of the error terms was assessed and verified.

Table 6: Normality Test for Third Model

W	**p-value**	**α**
0.99944	0.7384	0.05

H_0: Data is normal vs. H_a: Data is not normal

From the results of the normality test in Table 6, it is known that the p-value of the third model is 0.7384 which is greater than $\alpha = 0.05$. Hence, the model satisfies the normality of error terms.

c. Homoscedasticity

The homoscedasticity of the model was assessed through the Breusch-Pagan Test. The output of the test is exhibited in Table 7.

Table 7: Homoscedasticity Test for Third Model

Chi-square	**Degrees of freedom**	**p-value**	**α**
3.957957	1	0.04665	0.05

H_0: Homoscedasticity among residuals vs. H_a: Heteroscedasticity among residuals

The test results reveal a p-value of 0.04665, which is less than $\alpha = 0.05$. This leads to the rejection of the null hypothesis of homoscedasticity, suggesting that the residual variances are not constant or heteroscedastic. Consequently, variable transformation is necessary. Among the four transformation methods considered, only the reciprocal transformation proved effective; however, it cannot be applied due to the categorical nature of the predictor variables. This limitation necessitated the removal of the variables *Volunteering* and *Extracurricular*, prompting the formation of a revised model given by

$$GPA = \beta_0 + \beta_1 Absences + \beta_2 Age + \beta_3 Gender + \beta_4 Music + \beta_5 Parental\ SupportLow + \beta_6 ParentalSupportModerate + \beta_7 ParentalSupportNone + \beta_8 ParentalSupportVeryHigh + \beta_9 Sports + \beta_{10} logStudyTime + \beta_{11} Tutor$$

Table 7: Parameter Estimates of the Fourth Model

	Estimate	Std. Error	t-value	p-value
(Intercept)	3.2171508	0.0728242	44.177	< 2e-16 ***
Absences	-0.0994126	0.0005639	-176.291	< 2e-16 ***
Age	-0.0091071	0.0042602	-2.138	0.0326 *
GenderMale	-0.0199998	0.0095568	-2.093	0.0365 *
MusicYes	0.1405012	0.0120099	11.699	< 2e-16 ***
ParentalSupportLow	-0.2934916	0.0137698	-21.314	< 2e-16 ***
ParentalSupportModerate	-0.1602479	0.0123239	-13.003	< 2e-16 ***
ParentalSupportNone	-0.4682912	0.0183064	-25.581	< 2e-16 ***
ParentalSupportVery High	0.1552779	0.0171043	9.078	< 2e-16 ***
SportsYes	0.1927320	0.0103937	18.543	< 2e-16 ***
logStudyTime	0.1353542	0.0046909	28.855	< 2e-16 ***
TutorYes	0.2519196	0.0104024	24.217	< 2e-16 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 0.2331 on 2380 degrees of freedom				
Multiple R-squared: 0.9354, Adjusted R-squared: 0.9351				
F-statistic: 3133 on 11 and 2380 DF, p-value: < 2.2e-16				

The normality of the fourth model was asserted through the Shapiro-Wilk normality test, which produced a p-value of 0.5492, which is greater than the significance level

$\alpha = 0.05$. Thus, the null hypothesis that the data is normal cannot be rejected as manifested in Table 8.

Table 8: Normality Test for the Fourth Model

W	p-value	α
0.99932	0.5492	0.05

H_0: Data is normal vs. H_a: Data is not normal

Now, we check the homoscedasticity of the fourth model through its residuals plot shown below in Figure 3.

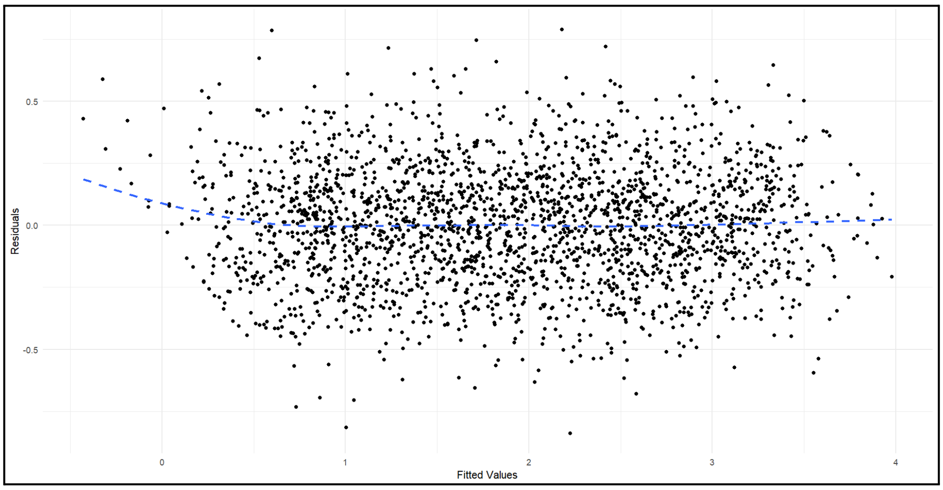

Figure 3: Residuals Plot of the Fourth Model

The residuals versus fitted values plot shows a random scatter of residuals around the horizontal line at zero, with no discernible funnel shape or pattern, indicating that the variance of the residuals is relatively consistent across all levels of fitted values. This suggests that the assumption of homoscedasticity is met. Additionally, the absence of increasing or decreasing spread supports the conclusion that the model does not exhibit heteroscedasticity. However, this visual assessment should be complemented by statistical tests, such as the Breusch-Pagan test, to confirm these observations. Following this, the homoscedasticity of the revised model was validated using the Breuch-Pagan Test. The results of this test are presented in Table 9.

Table 9: Homoscedasticity Test for the Fourth Model

Chi-square	Degrees of freedom	p-value	α
2.006917	1	0.15658	0.05

H_0: Homoscedasticity among residuals vs. H_a: Heteroscedasticity among residuals

Based on the test results, the p-value was calculated to be 0.15658, which exceeds the threshold of 0.05. Therefore, the residual variances are assumed to be constant, indicating homoscedasticity.

d. Independence of error terms

The Durbin-Watson test was conducted to evaluate the autocorrelation of error terms, and the results are presented below.

Table 10: Autocorrelation Test of Error Terms for the Fourth Model

Autocorrelation	Durbin-Watson Statistic	p-value
0.03262639	1.934572	0.096

H_0: No autocorrelation among residuals vs. H_a: Residuals are autocorrelated

The results show that the autocorrelation between residuals and their lagged values is close to zero, suggesting weak autocorrelation. Furthermore, the Durbin-Watson statistic is approximately 2, indicating the absence of autocorrelation among the residuals. Thus, the independence of error terms is upheld.

e. No Multicollinearity

The Variation Inflation Factor (VIF) test was conducted to evaluate the multicollinearity among the independent variables. The output of the test is shown in Table 11.

Table 11: Multicollinearity Test Result for the Fourth Model

	GVIF	DF	Adjusted GVIF [GVIF(1/(2*Df))]
Absences	1.003006	1	1.001502
Age	1.008338	1	1.004160

Gender	1.004423	1	1.002209
Music	1.003831	1	1.001914
ParentalSupport	1.011359	4	1.001413
Sports	1.005057	1	1.002526
LogStudyTime	1.003350	1	1.001673
Tutor	1.002811	1	1.001404

The table presents the multicollinearity test results for the fourth model using GVIF (Generalized Variance Inflation Factor). All GVIF values are close to 1, indicating low multicollinearity among the predictors. Moreover, the adjusted GVIF values, which account for degrees of freedom, are also close to 1, further confirming the absence of significant multicollinearity. This shows that the variance of the regression coefficients is not inflated due to the relationships between predictors.

f. Presence of Outliers

The presence of outliers was determined by using Cook's distance, and the results are presented in Figure 4.

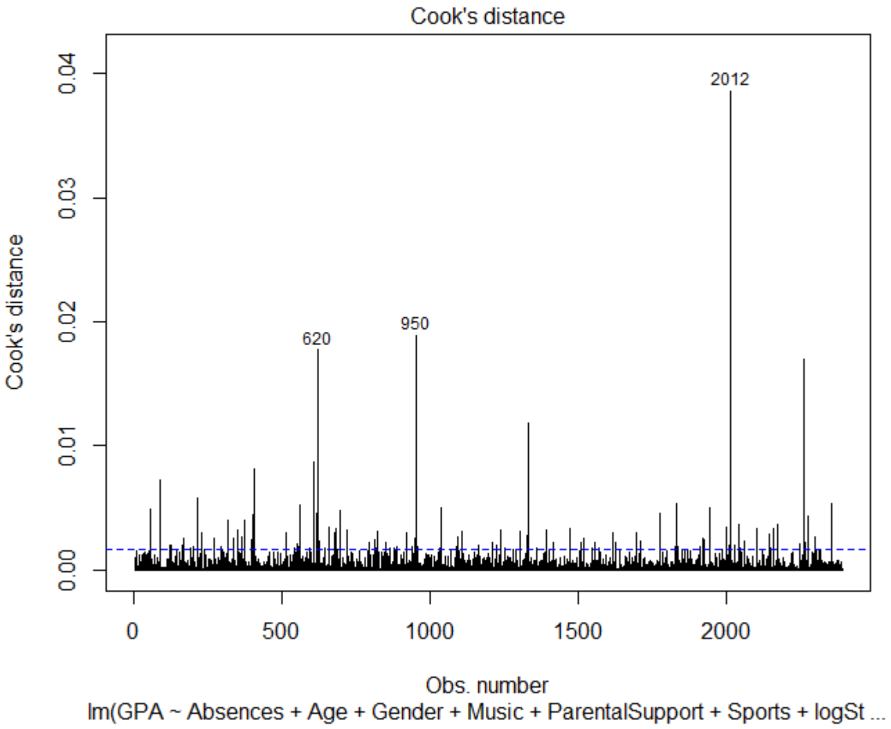

Figure 4: Cook's distance for the Fourth Model

Figure 4 reveals that observations 620, 950, and especially 2012 are significantly above the threshold line, indicating potential outliers based on Cook's distance values. The cutoff for Cook’s Distance was calculated as $\frac{4}{n-k-1}$, where $n = 2392$ is the total number of observations and $k = 8$ is the number of predictors in the model. This yielded a cutoff value of approximately 0.00168. While observation 950 exceeded this threshold, it was retained as its Cook's distance had no significant effect on the R^2 value, and was considered theoretically relevant. In contrast, observations 620 and 2012 were identified as influential outliers and were subsequently removed. Observation 620 reported 14.54 hours of study time per week with a GPA of 2.23, while observation 2012 reported 14.57 hours of study time per week with a GPA of 2.50. These cases were unusual as high study time was paired with relatively low academic performance. Additionally, both observations recorded very high parental support but had 15 and 14 absences, respectively, further deviating from expected patterns. Because of this, both observations were removed. This adjustment resulted in the final model, where all assumptions were satisfied. The parameter estimates for the final model are shown in Table 12.

Table 12: Parameter Estimates for the Final Model

	Estimate	Std. Error	t-value	p-value
(Intercept)	3.1996741	0.0725979	44.074	< 2e-16 ***
new_data$Absences	-0.0993737	0.0005616	-176.962	< 2e-16 ***
new_data$Age	-0.0088240	0.0042419	-2.080	0.0376 *
new_data$GenderMale	-0.0200187	0.0095158	-2.104	0.0355 *
new_data$MusicYes	0.1412340	0.0119552	11.814	< 2e-16 ***
new_data$ParentalSupportLow	-0.2904947	0.0137196	-21.174	< 2e-16 ***
new_data$ParentalSupportModerate	-0.1576120	0.0122785	-12.836	< 2e-16 ***
new_data$ParentalS	-0.4657962	0.0182287	-25.553	< 2e-16 ***

upportNone				
new_data$ParentalSupportVery High	0.1575766	0.0170315	9.252	< 2e-16 ***
new_data$SportsYes	0.1920664	0.0103515	18.554	< 2e-16 ***
new_data$logStudyTime	0.1404468	0.0047859	29.346	< 2e-16 ***
new_data$TutorYes	0.2508866	0.0103614	24.214	< 2e-16 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 0.2321 on 2378 degrees of freedom				
Multiple R-squared: 0.936, Adjusted R-squared: 0.9357				
F-statistic: 3162 on 11 and 2378 DF, p-value: < 2.2e-16				

Following diagnostic checks and validation of assumptions, the estimated model is defined as follows:

$$\widehat{GPA} = 3.1996741 - (0.0993737)\,Absences - (0.0088240)Age - (0.0200187)\,Gender + (0.1412340)\,Music - (0.2904947)\,Parental\,SupportLow - (0.1576120)\,ParentalSupportModerate - (0.4657962)\,ParentalSupportNone + (0.1575766)\,ParentalSupportVeryHigh + (0.1920664)\,Sports + (0.1404468)\,logStudyTime + (0.2508866)\,Tutor$$

This demonstrates that the final model for the multiple regression analysis provides a comprehensive understanding of the factors influencing the dependent variable. All predictors in the model, except Age and Gender, show highly significant effects on the outcome, as indicated by their extremely small p-values. The coefficient estimates highlight the direction and magnitude of these effects, with, for instance, Absences having a negative impact on the dependent variable, while variables like Sports and Tutoring have positive effects. Furthermore, the model exhibits a strong fit, with an R^2 value of 0.936, implying that 93.6% of the variation in the dependent variable is explained by the predictor variables included in the final model, while the remaining percentage is attributed to factors outside the scope of the regression model. Additionally, the F-statistic of 3162 and its associated p-value ($< 2.2e - 16$) confirm that the model as a whole is statistically significant. Overall, this final model effectively captures the relationships between the predictors and the dependent variable.

V. CONCLUSIONS AND RECOMMENDATIONS

The final model includes eight explanatory variables, namely, the number of absences (Absences), age of student (Age), gender (Gender), music participation (Music), level of parental support (ParentalSupport), participation in sports (Sports), log transform of study time (logStudyTime), and tutoring (Tutor), which are all significant at the 0.05 level. The model has an R^2 value of 0.936 and an adjusted R^2 value of 0.9357.

In the final model, student GPA is inversely proportional to the number of absences, which suggests that regular attendance and participation in class greatly impact their academic performance. Similarly, student GPA is also inversely proportional to age and gender. This might be because older students have more responsibilities, thus having less time for studying. However, since age has a relatively small coefficient, it does not strongly affect student GPA in this model. As for gender, the sample used suggests that male students tend to have lower GPAs compared to female students. Since gender also has a small coefficient, its impact on student GPA is minimal.

Moreover, the model shows that student GPA is also inversely proportional to low, moderate, and no parental support given to students. It suggests that parental support is important and that supportive environments help give students motivation to engage more in their studies. Students who received academic help at home, encouragement from parents, and positive reinforcement tend to have better academic performance. This is reflected in the directly proportional relationship between GPA and very high parental support.

It is also shown that students who participated in music and sports also received a higher GPA. This suggests that participation in sports and music may help in student discipline and time management. Additionally, studies show that participation in these activities helps boost mental health among students, which may in turn affect their overall GPA. Aside from these, the model also shows that student GPA is directly proportional to the log of study time and tutoring. This aligns with previous studies showing that the amount of study time weekly and tutoring have a significant effect on academic performance. Having an increased study time allows students to revisit concepts and master skills such as problem-solving and critical thinking. Moreover, tutoring may also address specific gaps in learning and strengthen areas that need improvement.

Overall, the final model is acceptable since it aligns with previous studies about how factors such as study time, parental support, participation in class, and extracurricular activities influence student performance. The results from the study only differ with how age affects student GPA as previous studies show that older students may outperform younger students. A possible reason for this is that older students may have additional responsibilities compared to younger students. Also, it is important to note the samples might have differed from the samples used in the previous studies.

There are several studies regarding the factors affecting student GPA, hence, we cannot say that we have included all possible indicators. To improve the model, future researchers may include more indicators, such as socioeconomic backgrounds, peer influence, intelligence quotient (IQ), and more. Also, since the data used in this study focuses on students from 15 to 18 years old, it might also be beneficial to explore more age brackets to gain insights into how other factors affecting student GPA evolve over time.

VI. References

Aissaoui, O. E., Madani, Y. E. a. E., Oughdir, L., Dakkak, A., & Allioui, Y. E. (2020). A multiple linear Regression-Based approach to predict student performance. *In Advances in intelligent systems and computing*, 9–23. https://doi.org/10.1007/978-3-030-36653-7_2

Ali, S., Haider, Z., Munir, F., Khan, H., & Ahmed, A. (2013). Factors contributing to the students academic performance: A case study of Islamia University Sub-Campus. *American Journal of Educational Research, 1*(8), 283–289. https://doi.org/10.12691/education-1-8-3

Andrietti, V., & Velasco, C. (2015). Lecture attendance, study time, and academic performance: a panel data study. *The Journal of Economic Education, 46*(3), 239–259. https://doi.org/10.1080/00220485.2015.1040182

Bahali, K., Tahiroğlu, A., & Avci, A. (2009). The clinical features of children and adolescents with school refusal. *Anatolian Journal of Psychiatry, 10*, 310-317

Brew, E. A., Nketiah, B., & Koranteng, R. (2021). A Literature Review of Academic Performance, an Insight into Factors and their Influences on Academic Outcomes of Students at Senior High Schools. *OALib, 08*(06), 1–14. https://doi.org/10.4236/oalib.1107423

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary educational psychology, 27*(2), 270-295.

Çelikkol, Ö. (2023). Impacts of the school starting age on academic achievement.

Chemers, M. M., Hu, L.-t., & Garcia, B. F. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology, 93*(1), 55–64. https://doi.org/10.1037/0022-0663.93.1.55

Çimen, S. (2000). Investigation of Psycho-Socia Development Attributes of Five Six Year Kindergarten Children Attending to Ankara Universtiy Kindergartens. Ankara University.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*(1), 13-21.

Dugan, B. D., Kyle, J. A., Kyle, C. W., Birnie, C., & Wahba, W. (2011). Integrating spirituality in patient care: preparing students for the challenges ahead. *Currents in Pharmacy Teaching and Learning, 3*(4), 260–266. doi:10.1016/j.cptl.2011.07.004

Fan, X., & Chen, M. (2001). Parental involvement and students' academic achievement: A meta-analysis. *Educational psychology review, 13*, 1-22.

Hardy, I. (2014). “I”m just a numbers person’: the complexity, nature and effects of the quantification of education. *International Studies in Sociology of Education, 25*(1), 20–37. doi:10.1080/09620214.2014.972971

Kuh, G. D., Kinzie, J., Buckley, J. A., Bridges, B. K., Hayek J. C. (2006). What Matters to Student Success: A Review of the Literature. *Commissioned Report for the National Symposium on Postsecondary Student Success: Spearheading a Dialog on Student Success, 8*.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2007). A Meta-Analysis of the Predictive Validity of the Graduate Management Admission Test (GMAT) and Undergraduate Grade Point Average (UGPA) for Graduate Student Academic Performance. *Academy of Management Learning & Education, 6*(1), 51–68. doi:10.5465/amle.2007.24401702

La Paro, K.M. & Pianta, R.C. (2000). Predicting children‟s competence in early school years: A meta-analytic review. *Review of Educational Research, 70*(4), 443-484. 10.3102/00346543070004443

Mansell W., James M. & the Assessment Reform Group. (2009). Assessment in schools. Fit for purpose? A Commentary by the Teaching and Learning Research Programme. *London: Economic and Social Research Council, Teaching and Learning Research Programme.*

Niu, L. (2018). A review of the application of logistic regression in educational research: common issues, implications, and suggestions. *Educational Review, 72*(1), 41–67. https://doi.org/10.1080/00131911.2018.1483892

OECD Policy Briefs. (2005). Formative assessment: Improving learning in secondary classrooms. *Paris.* Retrieved from http://www.oecd.org/dataoecd/19/31/35661078.pdf on June 11, 2011.

Ramey, S. L., Lanzi, R. G., Phillips, M. M., & Ramey, C. T. (1998). Perspectives of former head start children and their parents on school and the transition to school. *The Elementary School Journal, 98*(4), 311–327. https://doi.org/10.1086/461898

Reardon, S. F. (2018). The widening academic achievement gap between the rich and the poor. *In Social stratification*, 536-550.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological bulletin, 138*(2), 353.

Shafiq, F., Siddiquah, A. (2011). Effect of Classroom Quizzes on Graduate Students' Performance. *International Journal of Academic Research,* 3(5), 76-79.

Shanlax International Journal of Education, 11(S1-July), 208–215. https://doi.org/10.34293/education.v11is1-july.6186

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research, 75*(3), 417-453.

Sternberg, R.J., Grigorenko, E., & Bundy, D.A. (2001). The Predictive Value of IQ. *Merrill-Palmer Quarterly 47*(1), 1-41. https://dx.doi.org/10.1353/mpq.2001.0005.

Tadese, M., Yeshaneh, A., & Mulu, G. B. (2022). Determinants of good academic performance among university students in Ethiopia: a cross-sectional study. *BMC Medical Education*, *22*(1). https://doi.org/10.1186/s12909-022-03461-0

Van Dusen, B., & Nissen, J. (2019). Modernizing use of regression models in physics education research: A review of hierarchical linear modeling. *Physical Review Physics Education Research, 15*(2), 020108.

Voyles, M. J. (2011). Student academic success as related to student age and gender.

Yang, F., Li, F. W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers & Education, 123*, 97-108. https://doi.org/10.1016/j.compedu.2018.04.006.

Yang, S. J., Lu, O. H., Huang, A. Y., Huang, J. C., Ogata, H., & Lin, A. J. (2018). Predicting students' academic performance using multiple linear regression and principal component analysis. *Journal of Information Processing, 26*, 170-176.