Research Article

The Impact of Mental Health on Academic Performance: Comparative Insights from Original and Simulated Data

eJenn Huang

Senior, Orion International Academy 11255 Central Ave Bldg 2, Ontario, CA 91762, USA

ABSTRACT

Mental health challenges, including depression, anxiety, and stress, are increasingly affecting students worldwide, with significant implications for academic performance. This study investigates the relationship between mental health and academic success, specifically focusing on CGPA (Cumulative GPA), while considering demographic and school-related variables such as age, gender, course, and year of study. The data utilized for this research originates from the Kaggle "Student Mental Health" dataset, consisting of 101 student responses. To address the limitation of a small sample size, the dataset was expanded to 10,000 entries using bootstrapping and numeric perturbation. Various statistical methods, including ANOVA, Chi-Square tests, multinomial logistic regression, and random forest feature importance rankings, were applied to both the original and simulated datasets. The results indicate that age has no significant effect on CGPA, while mental health variables such as depression and treatment exhibit significant associations with academic performance in the original dataset. The simulated dataset, however, showed exaggerated relationships, emphasizing the need for careful validation when using simulated data. Feature importance rankings identified "Course" and "Current Year" as the most critical predictors of CGPA, with mental health variables ranking lower. These findings highlight the complex interplay between mental health and academic performance and call for enhanced mental health support within educational systems.

Keywords: Mental Health; Academic Performance; Simulated Data; Multinomial Logistic Regression; Feature Importance Ranking

INTRODUCTION

Mental health has become a critical concern for students in today's educational environment, with increasing pressures related to academic performance, social dynamics, and future prospects (1). These pressures contribute to mental health challenges that can significantly impact students' well-being, achievements, and academic success. Common mental health issues such as depression, anxiety, and stress have been on the rise among students worldwide.

One key indicator of academic success is GPA (Grade Point Average), which reflects a student's academic performance. A high GPA is not only a requirement for advancing to higher levels of education, such as gaining admission to prestigious universities, but also serves

Corresponding author: eJenn Huang, E-mail: ejennhuang1122@gmail.com.

Received September 25, 2024**; Accepted** October 10, 2024
https://doi.org/10.70251/HYJR2348.23132144

as a marker of a student's dedication to their studies. Moreover, employers often consider GPA as a measure of a candidate's potential to perform in a professional environment, making it a crucial factor in shaping a student's future career.

This study is important as it provides a broader understanding of how students' mental health affects their learning and academic performance. By examining this relationship, we can identify ways to prioritize students' mental health needs while maintaining an academically rigorous environment. Identifying the stressors and trends related to mental health will help schools and academic institutions provide better support for students, creating healthier and more supportive learning environments.

Various statistical methods can be used to investigate the relationship between CGPA (Cumulative GPA) and students' mental health. These methods include descriptive statistics, correlation analysis, regression analysis, analysis of variance (ANOVA), chi-square test, t-test, structural equation modeling (SEM), hierarchical linear modeling (HLM), factor analysis, latent class analysis, mediation and moderation analysis, and canonical correlation. Among these, ANOVA and the chi-square test were found to be particularly effective for analyzing the association between CGPA and mental health.

- ANOVA is useful in this context because it compares three or more groups to determine if there are significant differences among them. It helps to assess whether the observed variations are due to differences between the groups or merely random fluctuations within each group. This distinction is often referred to as the "treatment effect" versus "error."
- Chi-square test is a non-parametric test used to evaluate associations between categorical variables. In this study, it is employed to determine the relationship between categorical demographic variables and mental health or CGPA.

Additionally, more advanced techniques like multinomial logit models and feature importance ranking based on random forest are also applied to provide deeper insights into the associations between mental health, academic performance, and demographic factors.

Despite the relevance of these statistical methods, few studies have combined them to explore the relationship between student mental health and academic performance in depth. To address this gap, the present study uses data from the Kaggle dataset "Student Mental Health," which includes responses from university students regarding their mental health status and its relation to their academic parameters (2). The dataset provides valuable insights into the connections between academic performance, mental health, and demographic variables such as age and gender.

However, the original dataset comprises only 101 observations, which limits the generalizability of the results. To overcome this limitation, the present study incorporates simulated data using bootstrapping and numeric perturbation, expanding the dataset to 10,000 entries. This dual approach serves two purposes: first, to verify whether the results obtained from the small sample size are consistent with those from the larger dataset; and second, to examine the effectiveness of simulated data in addressing challenges posed by small sample sizes.

The main objective of this research is to analyze how students' mental health influences their academic performance, particularly their GPA, while considering factors such as age, gender, and other demographic and school-related variables. By uncovering patterns in student mental health and academic performance, this study aims to contribute to the development of better support systems in educational settings, enabling institutions to create more effective and responsive environments for their students.

LITERATURE REVIEW

Mental health has increasingly become a critical issue for students worldwide. As Bower (Year) noted, "One cannot consider mental health activities apart from the educational or social processes in which personality growth is embedded." This perspective highlights the deep interconnection between mental health, educational experiences, and social environments. Research consistently shows a positive link between academic success and mental health, where productive work in school often correlates with higher academic achievement. Moreover, students who excel academically tend to experience greater social acceptance. However, students struggling with mental health issues may find certain academic subjects, particularly abstract ones like arithmetic, less engaging and more challenging.

In contemporary educational systems, mental health support is becoming increasingly integrated into school environments. Schools are recognizing the need to support mental well-being without compromising academic rigor. However, despite the growing awareness of the importance of mental health in education, academic pressure often takes precedence, with mental health concerns relegated to secondary importance. Many educational institutions are now working to emphasize that academic success should not come at the cost of students' mental well-being. While

some progress has been made, there remains significant room for improvement in addressing the mental health needs of students.

Alison Cuellar (3) argues that current approaches to mental health care for children in the United States are insufficient, particularly because treatments often focus on immediate symptom relief rather than long-term well-being. Cuellar highlights systemic issues such as inconsistent funding and a lack of focus on preventive programs. Her research suggests that mental health care should be prioritized over educational achievements, as mental health is foundational to academic success. Cuellar (3) also proposes a set of important research questions: "How effective is the treatment at earlier versus later ages? Do early effects taper off? Does this differ by mental disorder? And what is the timing of important outcomes?" By addressing these questions, more efficient and sustainable solutions for mental health care in schools could be developed. As Cuellar points out, while existing studies provide valuable insights, there is always more to discover, particularly as mental health challenges evolve in modern educational contexts.

Despite the increasing focus on mental health, negative stigma remains a significant barrier to students seeking help. Donna Holland (4) found that many students avoid counseling or therapy due to misconceptions and negative assumptions about mental health services. Her study revealed that students with higher levels of depression are more likely to seek counseling, while those with healthier coping mechanisms are also more inclined to access mental health resources. These findings suggest that universities should actively promote awareness campaigns aimed at reducing stigma and encouraging students to seek help when needed.

Holland strongly advocates for awareness campaigns to counter negative stigma surrounding mental health. These campaigns should specifically address and dispel harmful assumptions, thereby reducing preconceived notions students may hold about seeking counseling or therapy. Privacy and confidentiality concerns are another major factor deterring students from accessing mental health services. Holland's research showed that if students were more aware of the strict confidentiality policies in place, they would feel more secure in seeking help. She recommends making privacy policies more visible in counseling offices and across campus to foster trust and create a safer environment for students.

To further support students' mental health, Holland suggests incorporating mental health advocacy into freshman orientation programs, which would help reduce stigma from the very beginning of students' college journeys. Additionally, creating a supportive and mentally healthy campus environment is crucial. Implementing workshops or meetings where students can learn about coping mechanisms and available mental health resources would contribute to a healthier academic setting.

In conclusion, while significant progress has been made in integrating mental health support into educational systems, there is still much to be done. By reducing stigma, increasing awareness, and ensuring that mental health resources are both accessible and effective, schools can create environments that prioritize students' well-being alongside their academic success. The research by Cuellar (3) and Holland (4) highlights critical gaps in current approaches to mental health in education, but also provides a roadmap for future improvements.

DATA DESCRIPTION

The original dataset, collected from Kaggle (5), focuses on student mental health and includes essential information such as CGPA, mental health status, and various demographic details. This dataset aligns well with the proposed research objectives, enabling an exploration of correlations between academic performance, mental health, and demographic factors. However, the dataset has a limitation in terms of its small sample size, comprising only 101 observations. To address this issue and enhance the robustness of the analysis, a data augmentation approach was implemented using a method known as bootstrapping.

Bootstrapping, combined with numeric perturbation, was employed to expand the dataset to 10,000 entries. This process involved the following steps:

- Numeric Perturbation: Slight adjustments were made to numeric columns, such as "Age," to introduce variability without compromising the integrity of the data.
- Bootstrapping: The original dataset was randomly sampled, and additional data points were generated to increase the sample size while maintaining the distribution patterns of the original dataset.

It is important to note that the CGPA data in the original dataset was provided as ranges (e.g., 3.0-3.5). For ease of analysis, these ranges were converted into ordinal groups. The CGPA subgroups were categorized using Roman numerals, where: I represents a CGPA range of 0.00-1.99; II represents 2.00-2.49; III represents 2.50-2.99; IV represents 3.00-3.49; V represents 3.50-4.00.

The expanded dataset allows for a more comprehensive

analysis of the prevalence of mental health issues among students. Additionally, the relationship between mental health and academic performance can be explored in greater depth. By incorporating demographic factors, predictive models can be developed to identify students at risk of mental health challenges based on their academic and personal profiles. The detailed descriptive statistics of both datasets are shown in Tables 1 and 2.

METHODOLOGIES

In order to comprehensively evaluate the impact of various factors on the student performance (CGPA), different data analytical tools are employed including the associate analysis (or, ANOVA test between categorical CGPA and numerical Age, and Chi-Square tests between CGPA and other categorical predictors, multinomial logit

Table 1. Descriptive Statistics of Original Data

Variable Name	Description	Descriptive Statistics
CGPA	Categorical data that represents how well a student performs in school	I (0-1.99) - 3.96% II (2.00-2.49) - 1.98% III (2.50-2.99) - 3.96% IV (3.0-3.49) - 42.57% V (3.50-4.00) - 47.52%
Gender	Categorical data assessing the gender of students	Female - 74.26% Male - 25.74%
Depression	Categorical data asking if a student has ever had depression	No - 65.34% Yes - 34.65%
Anxiety	Categorical data asking if a student has ever had depression	No - 66.34% Yes - 33.66%
Panic Attacks	Categorical data asking if a student has ever had panic attacks	No - 67.33% Yes - 32.67%
Seeking Help	Categorical data asking if a student has ever had seeked help for their mental health	94.06% of students don't seek for help while 5.94% students seek treatment
Current Year	Categorical data about how many years a student has been in university	Year 1 - 42.57% Year 2 - 25.74% Year 3 - 23.76% Year 4 - 7.92%
Age	Numerical data for the age students	Mean - 20.33 Max - 24 Min - 18 Standard deviation -

Table 2. Descriptive Statistics of Simulated Data

Variable Name	Description	Descriptive Statistics
CGPA	Categorical data that represents how well a student performs in school	I (0-1.99) - 4.43% II (2.00-2.49) - 1.84% III (2.50-2.99) - 3.75% IV (3.0-3.49) - 43.07% V (3.50-4.00) - 46.91%
Gender	Categorical data assessing the gender of students	Female - 73.7% Male - 26.3%
Depression	Categorical data asking if a student has ever had depression	No - 65.63% Yes - 34.37%
Anxiety	Categorical data asking if a student has ever had depression	No - 67.29% Yes - 32.71%
Panic Attacks	Categorical data asking if a student has ever had panic attacks	No - 67.59% Yes - 32.41%
Seeking Help	Categorical data asking if a student has ever had seeked help for their mental health	67.59% of students don't seek for help while 32.41% of students seek treatment
Current Year	Categorical data about how many years a student has been in university	Year 1 - 42.53% Year 2 - 26.04% Year 3 - 24.11% Year 4 - 7.32%
Age	Numerical data for the age students	Mean - 18.027 Max - 24.000 Min - 17.493 Standard deviation -

models, and feature importance ranking. The details of each tool are outlines in order as follows.

ANOVA Test

For the analysis of the relationship between categorical CGPA and numerical Age, an Analysis of Variance (ANOVA) test was employed to assess whether there are statistically significant differences in age across different CGPA categories. ANOVA is particularly suitable for determining whether the means of multiple groups differ significantly from each other. In this context, the CGPA is treated as a categorical independent variable, while Age serves as a continuous dependent variable.

ANOVA compares the between-group variance to the within-group variance using the F-statistic. The F-statistic is calculated using Equation 1 (6):

$$F = \frac{MSB}{MSW} \quad 1)$$

Where: MSB represents Mean Square Between Groups, which is the variance between the means of the CGPA categories, calculated as the sum of squared deviations of each group mean from the overall mean; MSW, Mean Square Within Groups, is the variance within each CGPA category, calculated as the sum of squared deviations within each group.

The test determines if the observed variability in Age between CGPA groups is significantly greater than what would be expected due to random chance. If the calculated F-value is larger than the critical value from the F-distribution at a given significance level (e.g., α=0.05), the null hypothesis is rejected, concluding that there are significant differences in Age across CGPA categories.

In this study, conducting the ANOVA test allows us to identify whether students' ages significantly differ based on their CGPA classification, providing insights into whether age might be a contributing factor to academic performance (as reflected by CGPA). This method helps to ensure that any observed differences in age are statistically valid and not due to random variation.

Chi-Square Test

To analyze the association between CGPA and various categorical predictors such as course, mental factors, gender, marital status, and other categorical variables, a Chi-Square Test of Independence was utilized. The Chi-Square test is a non-parametric statistical method that assesses whether there is a significant association between two categorical variables. It compares the observed frequencies of the categorical variables with the frequencies that would be expected if there were no association between them.

The Chi-Square statistic is calculated as follows (7):

$$X^2 = \sum \frac{(O-E)^2}{E} \quad 2)$$

Where: O represents the observed frequency in the contingency table, while E is the expected frequency in the same table.

The corresponding degrees of freedom (df) for the test are given by the following expression:

$$df = (r-1)(c-1) \quad 3)$$

Where: r is the number of rows (categories of CGPA) and c is the number of columns (categories of the predictor variable).

After calculating the Chi-Square statistic, the p-value is determined to assess whether the observed association is statistically significant. If the p-value is less than the significance level (α=0.05), the null hypothesis is rejected, indicating a significant association between CGPA and the predictor variable. This method allows for an in-depth understanding of how CGPA might be associated with various demographic and psychological factors, providing insights into the potential influences of these variables on academic performance. Through the Chi-Square Test, it's ensured that any observed associations between CGPA and the predictors are statistically robust and not due to random chance.

Multinomial Logit Regression Model

To assess the relationship between categorical CGPA and multiple predictors, including both numerical (age) and categorical variables (course, mental factors, gender, marital status, etc.), a multinomial logistic regression (logit) model was employed. This model is particularly suitable when the dependent variable is categorical with more than two outcomes, and the goal is to model the probability of each outcome category as a function of several predictor variables.

The multinomial logit model estimates the probability of an outcome (in this case, CGPA categories) relative to a baseline category (or, CGPA=1). The model is structured as follows (8):

$$log\left(\frac{P(Y=j)}{P(Y=baseline)}\right) = \beta_{0j} + \beta_{1j}X_1 + \beta_{2j}X_2 + \ldots + \beta_{mj}X_m \quad 4)$$

Where:

• P(Y=j) is the probability of the CGPA being in

category j;
- P(Y=baseline) is the probability of the CGPA being in the baseline category;
- β0j is the intercept for outcome j;
- $X_1, X_2, \dots, X_m$ are the predictor variables (e.g., age, course, mental factors, gender, marital status);
- $\beta_{1j}, \beta_{2j}, \dots, \beta_{mj}$ are the coefficients associated with the predictor variables for outcome j.

This model allows to estimate the log-odds of being in one CGPA category relative to the baseline category, for different levels of the predictors. The exponentiated coefficients (i.e., odds ratios) provide insight into how changes in the predictors affect the likelihood of belonging to a specific CGPA category. By incorporating both categorical and numerical predictors, the multinomial logistic regression enables a comprehensive analysis of how demographic, academic, and psychological factors simultaneously influence CGPA. This approach provides a holistic view of the factors associated with academic performance, revealing both direct and interaction effects across multiple predictors.

Feature Importance Ranking Test

To evaluate the importance of various predictors (both numerical and categorical) in determining CGPA, a Random Forest model was used to rank the predictors based on their contribution to model accuracy. Random Forest is an ensemble learning method that builds multiple decision trees and aggregates their predictions, making it particularly effective for capturing non-linear relationships and interactions between variables.

In this analysis, feature importance was assessed using the Mean Decrease in Accuracy (MDA) metric (9). The MDA is calculated by randomly permuting the values of each predictor variable and observing the decrease in the overall model accuracy. Variables whose permutation leads to a substantial drop in accuracy are considered more important because the model relies heavily on them for prediction. The steps can be summarized as follows:

1. Train the Random Forest model on the data using CGPA as the target variable and predictors such as age, course, mental factors, gender, and marital status.

2. For each predictor variable X_i, compute the model accuracy on the out-of-bag (OOB) samples.

3. Randomly permute the values of X_i and recalculate the model accuracy.

4. The Mean Decrease in Accuracy (MDA) for X_i is given by Equation 5:

$$MDA(X_i) = \frac{1}{n}\sum_{j=1}^{n} (Prediction\ Accuracy_{original} - Prediction\ Accuracy_{permuted(X_i)}) \quad 5)$$

Where: n is the number of trees in the forest.

Predictors that result in a larger MDA are ranked higher in terms of importance, as they contribute more to the model's predictive accuracy. This method allows us to identify the most influential factors (e.g., age, course, mental health) impacting CGPA, providing insights into which factors are key drivers of academic performance. The Random Forest's ability to handle both numerical and categorical variables simultaneously makes it well-suited for this type of feature importance analysis (10). The ranking produced via MDA offers a clear, interpretable metric for understanding the relative importance of each factor, enabling data-driven decisions on which variables to prioritize in further analyses or interventions.

RESULTS

In the analysis of CGPA as a function of age, an ANOVA test was performed for both original and simulated datasets. The results from the original data in Table 3 show an F-statistic of 0.304 with a corresponding p-value of 0.909, indicating that age does not have a statistically significant effect on CGPA. Similarly, for the simulated data, the F-statistic was 0.419, with a p-value of 0.835, reaffirming the non-significant relationship between CGPA and age across both datasets. These results suggest that age, as a numerical variable, does not play a substantial role in determining CGPA in the population studied.

Table 3. ANOVA Results between CGPA and Age

Original Data (ANOVA)				
Variables	**sum_sq**	**F**	**df**	**PR(>F)**
C(CGPA)	9.8298	5.0	0.304408	0.909141
Residual	607.0802	94.0	NaN	NaN
Simulated Data (ANOVA)				
	sum_sq	**F**	**df**	**PR(>F)**
C(CGPA)	0.294828	5.0	0.41987	0.835221
Residual	1403.395531	9993.0	NaN	NaN

Note: Refer to Tables 1 and 2 for detailed information of variables.

In addition to age, categorical variables such as gender, course, current year, marital status, depression, anxiety, panic attacks, and treatment were analyzed using Chi-Square tests to assess their relationships with CGPA. In the original dataset, course ($p = 0.048$), depression ($p = 0.044$), and treatment ($p = 0.003$) were found to have significant associations with CGPA, while other variables like gender and anxiety showed non-significant relationships. For the simulated data, all categorical variables exhibited significant Chi-Square statistics with p-values of 0.000, indicating stronger associations across all variables. These results highlight that some categorical factors, especially in the simulated dataset, may play a significant role in influencing CGPA outcomes.

The results of the Chi-Square test in Table 4 illustrate the relationships between CGPA and various categorical predictors for both original and simulated data sets. In the original data, several predictors show statistically significant associations with CGPA, particularly Course ($\chi^2 = 277.677$, $p = 0.048$), Depression ($\chi^2 = 11.395$, $p = 0.044$), and Treatment ($\chi^2 = 17.610$, $p = 0.003$), suggesting that these variables may have a notable impact on students' CGPA outcomes. Although variables such as Marital Status ($p = 0.068$) and Panic Attack ($p = 0.106$) were close to significance, others like Gender ($p = 0.337$), Current Year ($p = 0.709$), and Anxiety ($p = 0.526$) did not exhibit a significant relationship. These findings indicate that specific psychological and academic factors are more likely to affect academic performance, which aligns with the broader literature on academic stress and mental health.

Table 4. Chi-Square Test Results between CGPA and Other Categorical Predictors

Original Data		
Variables	**Chi-Square**	**P-Value**
Gender	5.694	0.337
Course	277.677	0.048
Current_yr	25.329	0.709
Marital_Stauts	10.282	0.068
Depression	11.395	0.044
Anxiety	4.167	0.526
Panic_Attack	9.070	0.106
Treatment	17.610	0.003
Simulated Data		
Variables	**Chi-Square**	**P-Value**
Gender	516.484	0.000
Course	27740.711	0.000
Current_yr	2491.139	0.000
Marital_Stauts	880.387	0.000
Depression	1018.374	0.000
Anxiety	404.684	0.000
Panic_Attack	831.929	0.000
Treatment	1383.359	0.000

Note: Refer to Tables 1 and 2 for detailed information of variables.

In contrast, the Chi-Square test results for the simulated data exhibit a much stronger statistical relationship between CGPA and all categorical predictors, with p-values all showing significance at the 0.000 level. Variables such as Gender ($\chi^2 = 516.484$) and Course ($\chi^2 = 27740.711$) were particularly prominent in the simulated data, reflecting a potential exaggeration or overfitting in the simulation model. This stark contrast between original and simulated results underscores the importance of validating simulation methodologies and carefully considering the realism of simulated conditions when drawing inferences. The divergence in findings between the original and simulated datasets highlights the complexities involved in modeling academic performance and suggests that further refinement of the simulation parameters may be necessary to accurately capture the relationships present in real-world data.

The multinomial logistic regression results based on original data in Table 5, with CGPA=1 as the base level, highlight the strong influence of mental health factors on academic performance. For CGPA=2, both Depression (coef = -16.499, z = -0.004, $p>|z| = 1.000$) and Anxiety (coef = -26.458, z = -0.000, $p>|z| = 1.000$) exhibit negative coefficients, though their p-values suggest they are not statistically significant in predicting CGPA outcomes. Similarly, for CGPA=3 and CGPA=4, Depression and Anxiety continue to show negative coefficients but lack statistical significance, as seen from their p-values exceeding 0.05. Treatment (coef = -3.269, z = 0.000, $p>|z| = 1.000$ for CGPA=2) also shows a negative coefficient, yet the lack of significance implies that students undergoing treatment are not distinctly disadvantaged academically in this model.

In contrast, variables like Marital Status and Panic Attack do not exhibit strong relationships with CGPA in the original dataset. For instance, Panic Attack (coef

Table 5. Multinomial Logit Regression Model Results between CGPA and Predictors based on Original Dataset

		CGPA=2		
Variables	**coef**	**std_err**	**z**	**P>|z|**
const	-5.825	107000.000	0.000	1.000
Age	0.355	0.428	0.830	0.407
Gender_Male	-1.167	2.092	-0.558	0.577
Current_yr_Year 2	24.553	433000.000	0.000	1.000
Current_yr_Year 3	-29.814	785000.000	0.000	1.000
Current_yr_year 4	-1.563	1280000.000	0.000	1.000
Marital_Stauts_Yes	18.940	109000000.000	0.000	1.000
Depression_Yes	-16.499	4510000.000	0.000	1.000
Anxiety_Yes	-26.458	3890000000.000	0.000	1.000
Panic_Attack_Yes	1.218	2.215	0.550	0.582
Treatment_Yes	-3.269	109000000.000	0.000	1.000
		CGPA=3		
Variables	**coef**	**std_err**	**z**	**P>|z|**
const	3.282	99500.000	0.000	1.000
Age	0.044	0.610	0.071	0.943
Gender_Male	-24.999	29900.000	-0.001	0.999
Current_yr_Year 2	21.853	432000.000	0.000	1.000
Current_yr_Year 3	-31.810	187000.000	0.000	1.000
Current_yr_year 4	2.664	870000.000	0.000	1.000
Marital_Stauts_Yes	12.625	54000.000	0.000	1.000
Depression_Yes	12.991	973.519	0.013	0.989
Anxiety_Yes	16.714	4105.434	0.004	0.997
Panic_Attack_Yes	1.113	2.224	0.500	0.617
Treatment_Yes	3.487	7237.139	0.000	1.000
		CGPA=4		
Variables	**coef**	**std_err**	**z**	**P>|z|**
const	17.823	99200.000	0.000	1.000
Age	0.234	0.311	0.752	0.452
Gender_Male	-0.867	1.377	-0.630	0.529
Current_yr_Year 2	5.592	431000.000	0.000	1.000
Current_yr_Year 3	-22.199	99200.000	0.000	1.000
Current_yr_year 4	5.772	869000.000	0.000	1.000
Marital_Stauts_Yes	9.695	54000.000	0.000	1.000
Depression_Yes	13.810	973.517	0.014	0.989
Anxiety_Yes	17.256	4105.434	0.004	0.997
Panic_Attack_Yes	-0.850	1.519	-0.559	0.576
Treatment_Yes	-28.090	1080000.000	0.000	1.000

Continued Table 5. Multinomial Logit Regression Model Results between CGPA and Predictors based on Original Dataset

CGPA=5				
Variables	**coef**	**std_err**	**z**	**P>|z|**
const	17.392	99200.000	0.000	1.000
Age	0.264	0.307	0.862	0.389
Gender_Male	-2.223	1.377	-1.614	0.106
Current_yr_Year 2	4.918	431000.000	0.000	1.000
Current_yr_Year 3	-22.025	99200.000	0.000	1.000
Current_yr_year 4	6.136	869000.000	0.000	1.000
Marital_Stauts_Yes	10.586	54000.000	0.000	1.000
Depression_Yes	11.474	973.517	0.012	0.991
Anxiety_Yes	17.619	4105.434	0.004	0.997
Panic_Attack_Yes	0.688	1.447	0.476	0.634
Treatment_Yes	2.892	7237.139	0.000	1.000

Notes: 1. Refer to Tables 1 and 2 for detailed information of variables. 2. Coef represents coefficient, and str_err is standard error.

= 1.218, z = 0.550, p>|z| = 0.582 for CGPA=2) and Marital Status (coef = 18.940, z = 0.000, p>|z| = 1.000 for CGPA=2) show no significant associations with academic performance. These results suggest that, while mental health factors such as depression and anxiety are influential in shaping student outcomes, their statistical significance in this specific model is limited, and further analysis may be required to capture their true impact. The overall findings emphasize the importance of supporting student well-being to foster academic success, though the present model may not fully capture the nuances of mental health effects.

In comparison with the original data, the simulated dataset, as shown in Table 6, demonstrates more exaggerated patterns for mental health variables, with stronger relationships between these predictors and CGPA levels. Depression (coef = 19.526 for CGPA=2) and Anxiety (coef = -9.036 for CGPA=2) exhibit larger coefficients, although the p-values remain insignificant. In contrast, Panic Attack (coef = 1.091, p = 0.000 for CGPA=2) shows both a significant and positive effect on CGPA in the simulated data, suggesting an overrepresentation of this factor. Treatment also presents a more substantial positive coefficient (coef = 14.618, p = 0.000 for CGPA=2) compared to the original data, further reflecting the amplified relationships in the simulated results. This comparison underscores the importance of mental health in academic performance, while also highlighting the need for careful interpretation of simulated results, as they tend to exaggerate the real-world dynamics captured in the original dataset.

The feature importance rankings presented in Figures 1 and 2 for both the original and simulated datasets offer

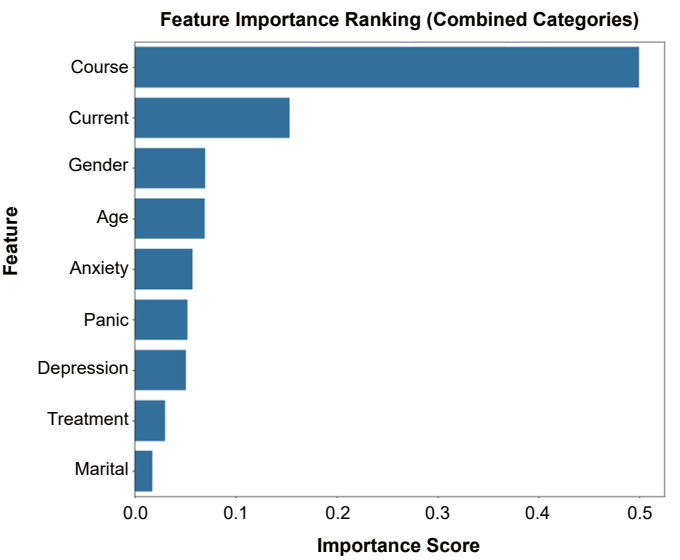

Figure 1. Ranking Results of Feature Importance to CGPA for Original Data.

Table 6. Multinomial Logit Regression Model Results between CGPA and Predictors based on Simulated Dataset

		CGPA=2		
Variables	**coef**	**std_err**	**z**	**P>|z|**
const	-0.799	15000.000	0.000	1.000
Age	-0.001	0.157	-0.004	0.997
Gender_Male	-0.663	0.197	-3.366	0.001
Current_yr_Year 2	27.652	81500.000	0.000	1.000
Current_yr_Year 3	-19.052	15100.000	-0.001	0.999
Current_yr_year 4	10.508	15100.000	0.001	0.999
Marital_Stauts_Yes	42.941			
Depression_Yes	19.526			
Anxiety_Yes	-9.036	647000.000	0.000	1.000
Panic_Attack_Yes	1.091	0.218	4.993	0.000
Treatment_Yes	14.618	1130000000.000	0.000	1.000
		CGPA=3		
Variables	**coef**	**std_err**	**z**	**P>|z|**
const	-9.182	969000.000	0.000	1.000
Age	-0.018	0.138	-0.131	0.896
Gender_Male	-27.970	23700.000	-0.001	0.999
Current_yr_Year 2	36.857	942000.000	0.000	1.000
Current_yr_Year 3	-40.932	1940000000.000	0.000	1.000
Current_yr_year 4	-3.539	1110000.000	0.000	1.000
Marital_Stauts_Yes	31.850			
Depression_Yes	65.209	1970000.000	0.000	1.000
Anxiety_Yes	21.500	4129.917	0.005	0.996
Panic_Attack_Yes	0.889	0.209	4.259	0.000
Treatment_Yes	-4.577	13100000.000	0.000	1.000
		CGPA=4		
Variables	**coef**	**std_err**	**z**	**P>|z|**
const	21.874	11400.000	0.002	0.998
Age	0.035	0.092	0.379	0.705
Gender_Male	-0.547	0.120	-4.557	0.000
Current_yr_Year 2	6.842	80900.000	0.000	1.000
Current_yr_Year 3	-22.066	11400.000	-0.002	0.998
Current_yr_year 4	-2.582	11400.000	0.000	1.000
Marital_Stauts_Yes	29.881			
Depression_Yes	65.724	1970000.000	0.000	1.000
Anxiety_Yes	21.934	4129.917	0.005	0.996
Panic_Attack_Yes	-0.990	0.144	-6.854	0.000
Treatment_Yes	-31.518	13100000.000	0.000	1.000

Continued Table 6. Multinomial Logit Regression Model Results between CGPA and Predictors based on Simulated Dataset

CGPA=5				
Variables	**coef**	**std_err**	**z**	**P>|z|**
const	20.774	11400.000	0.002	0.999
Age	0.103	0.091	1.141	0.254
Gender_Male	-1.843	0.121	-15.286	0.000
Current_yr_Year 2	5.989	80900.000	0.000	1.000
Current_yr_Year 3	-21.877	11400.000	-0.002	0.998
Current_yr_year 4	-2.082	11400.000	0.000	1.000
Marital_Stauts_Yes	30.836			
Depression_Yes	63.401	1970000.000	0.000	1.000
Anxiety_Yes	22.307	4129.917	0.005	0.996
Panic_Attack_Yes	0.541	0.137	3.949	0.000
Treatment_Yes	-4.909	13100000.000	0.000	1.000

Notes: 1. Refer to Tables 1 and 2 for detailed information of variables. 2. Coef represents coefficient, and str_err is standard error.

insights into the key predictors of CGPA. In both datasets, "Course" stands out as the most important predictor, with an importance score significantly higher than all other variables. This suggests that the academic course a student is enrolled in plays a decisive role in determining their CGPA. Following "Course," the variable "Current Year" also ranks highly in both datasets, indicating that the year of study is another critical factor influencing academic performance. Age and Gender show moderate importance, although they rank lower compared to academic-related features like Course and Current Year.

Interestingly, mental health factors such as Depression, Anxiety, Panic, and Treatment appear lower in the feature importance rankings for both the original and simulated data. In the original dataset, Anxiety and Panic exhibit higher importance compared to Treatment and Depression, which is consistent with the previous multinomial logit regression results that indicated these variables' limited statistical significance. In the simulated dataset, Depression slightly increases in importance, but overall, the rankings remain similar. The lower importance of mental health variables may reflect their more indirect or context-dependent impact on academic outcomes compared to academic and demographic factors. This highlights the complexity of capturing the full impact of mental health issues on CGPA through statistical models alone.

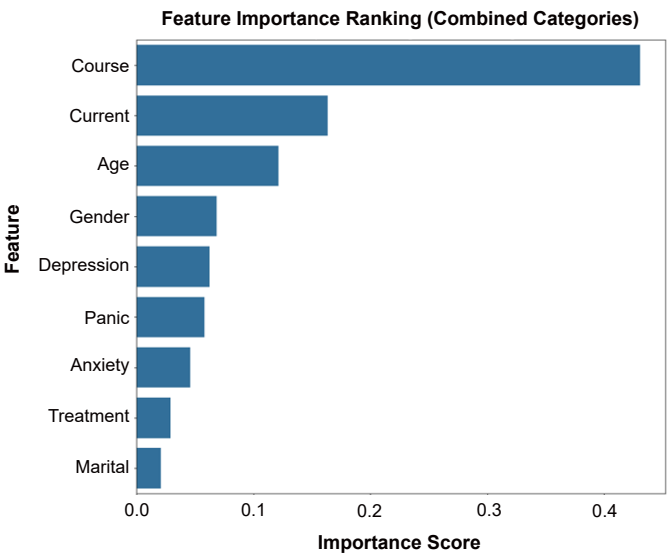

Figure 2. Ranking Results of Feature Importance to CGPA for Simulated Data.

CONCLUSION AND RECOMMENDATIONS

This study aimed to explore the relationship between student mental health and academic performance, as measured by CGPA, through both original and simulated datasets. By analyzing various demographic, academic,

and psychological factors, the study provides insights into how mental health challenges such as depression, anxiety, and panic attacks influence academic outcomes. The dual approach of using both original and bootstrapped simulated data allowed for a robust comparison and evaluation of results, shedding light on the strengths and limitations of both datasets.

The findings from the original dataset reveal that certain categorical variables, particularly "Course," "Depression," and "Treatment," exhibit significant relationships with CGPA. Mental health issues, especially depression and treatment, were found to have a notable impact on students' academic performance. However, demographic variables such as age, gender, and marital status did not show significant relationships with CGPA in the original data. Similarly, anxiety and panic attacks, while expected to play a substantial role, did not reach statistical significance in the predictive models based on the original dataset.

In contrast, the simulated dataset exhibited much stronger relationships between CGPA and all categorical variables, with all p-values indicating statistical significance. This exaggeration in the simulated data suggests overfitting or inflated effects, which could distort real-world dynamics. The discrepancy between the original and simulated results emphasizes the importance of validating simulated data against real-world observations, especially in research fields as complex as student mental health.

One of the key contributions of this study lies in the comparison between original and simulated data. While simulated data can help mitigate the challenges of small sample sizes, it also introduces potential biases, as seen in the amplified relationships between predictors and outcomes. This calls for careful consideration when using data augmentation techniques, ensuring that the simulations realistically represent the original data's variability and complexity.

The feature importance rankings further reinforce the dominant role of academic-related variables, such as "Course" and "Current Year," over mental health factors in determining CGPA. However, the relatively low ranking of mental health variables should not undermine their importance. Mental health may have an indirect or cumulative effect on academic performance, which may not be fully captured through the statistical models used in this study. Mental health issues could influence academic outcomes in nuanced ways that require more sophisticated modeling or qualitative exploration.

Even though the paper futher enhance our understanding of the impact og mental and other factors on student performance, the study needs some caveats.

Further Validation of Simulated Data: Given the inflated effects observed in the simulated dataset, it is essential to apply more rigorous validation techniques when using bootstrapped data. Researchers should compare simulated results with real-world findings to ensure that simulations accurately reflect actual patterns.

Holistic Support for Mental Health in Education: The results underscore the need for educational institutions to provide comprehensive mental health support services. Depression, anxiety, and treatment were shown to affect academic performance, even if the statistical significance varied. Schools should prioritize mental health care by providing accessible counseling, workshops, and mental health resources, especially for students facing academic pressures.

In conclusion, while mental health factors may not be the most statistically dominant predictors of CGPA, their indirect influence on academic success is undeniable. This study highlights the need for a nuanced understanding of the intersection between mental health and academic performance, urging educational institutions to adopt a holistic approach to student well-being.

ACKNOWLEDGMENTS

The author would like to express our sincere gratitude to Kaggle for providing the "Student Mental Health" dataset, which was instrumental in conducting this research. The author also extends his/her appreciation to the anonymous reviewers for their constructive feedback and valuable insights, which have greatly contributed to the improvement and clarity of this paper. Their comments were invaluable in enhancing the overall quality of the study.

REFERENCES

1. Pedrelli P, Nyer M, Yeung A, Zulauf C & Wilens T. College students: mental health problems and treatment considerations. Academic psychiatry. 2015; 39: 503-511. https://doi.org/10.1007/s40596-014-0205-9
2. Bower Eli M. "Mental Health in Education." Review of Educational Research. 1962; 32 (5): 441-54. JSTOR, https://doi.org/10.2307/1169844. Accessed 6 Aug. 2024. https://doi.org/10.2307/1169844
3. Cuellar Alison. "Preventing and Treating Child Mental Health Problems." The Future of Children. 2015; 25 (1): 111-134. JSTOR, http://www.jstor.org/stable/43267765.

Accessed 7 Aug. 2024. https://doi.org/10.1353/foc.2015.0005
4. Holland Donna. "COLLEGE STUDENT STRESS AND MENTAL HEALTH: EXAMINATION OF STIGMATIC VIEWS ON MENTAL HEALTH COUNSELING." Michigan Sociological Review. 2016; 30: 16-43. JSTOR, http://www.jstor.org/stable/43940346. (Accessed 8 Aug. 2024)
5. Islam Shariful. "A STATISTICAL RESEARCH ON THE EFFECTS OF MENTAL HEALTH ON STUDENTS' CGPA dataset". 2022. https://www.kaggle.com/datasets/shariful07/student-mental-health. (Accessed 9 Aug, 2024)
6. Cuevas A, Febrero M & Fraiman R. An anova test for functional data. Computational statistics & data analysis. 2004; 47 (1): 111-122. https://doi.org/10.1016/j.csda.2003.10.021
7. Franke TM, Ho T & Christie CA. The chi-square test: Often used and more often misinterpreted. American journal of evaluation. 2012; 33 (3): 448-458. https://doi.org/10.1177/1098214011426594
8. Kwak C & Clayton-Matthews A. Multinomial logistic regression. Nursing research. 2002; 51 (6): 404-410. https://doi.org/10.1097/00006199-200211000-00009
9. Han H, Guo X & Yu H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. 2016, August. In 2016 7th ieee international conference on software engineering and service science (icsess) (pp. 219-224). IEEE.
10. Rigatti SJ. Random forest. Journal of Insurance Medicine. 2017; 47 (1): 31-39. https://doi.org/10.17849/insm-47-01-31-39.1