IME 451 | Advanced Statistical Methods

Dr. Mohamed Gheith

Course Project

Predicting Student Academic Performance and Stress Level

By:

Maryana Kamal Megahed	**120210089**
Salma Tarek Salem	**120210125**
Marina Atef Wissa	**120210186**
Omar Mohamed Hussein	**120210288**
Logain Emad Eldin Awad	**120210340**

Table of Contents

1. Dataset Overview

Many students today are experiencing high levels of stress, which can significantly impact their academic performance, mental health, and overall well-being. Educators and counselors are keen to identify the key factors contributing to student stress to develop effective interventions and support systems. Understanding which students are at a higher risk of experiencing severe stress would enable decision-makers to design targeted programs, create a more supportive environment, and improve resources aimed at reducing stress levels and promoting mental wellness.

The dataset collected for this purpose provides a comprehensive view of various factors influencing student stress. Spanning psychological, physiological, environmental, academic, and social domains, the data captures the nuanced experiences of 1,100 students. Our goal is to analyze data using visualization and clustering to uncover factors contributing to stress, predict stress levels based on various factors, and subsequently predict academic performance by incorporating stress levels and academic factors, ultimately proposing measures to support students effectively.

2. Data Description

2.1 Key Classifications of Variables

To better understand the diverse factors contributing to student stress, the dataset can be divided into the following categories:

- **Psychological Factors**: These factors relate to the mental and emotional state of the students and include variables such as anxiety and depression. They capture students' internal psychological experiences, which are central to understanding their stress levels.
  - *anxiety_level*: Values range from 0 to 21, indicating varying levels of anxiety among students.
  - *self_esteem*: Values range from 0 to 30, representing the self-reported levels of self-esteem.
  - *mental_health_history*: Binary variable (0 for no history, 1 for history of mental health issues).
  - *depression*: Values range from 0 to 27, capturing the severity of depressive symptoms.
- **Physiological Factors**: These factors address the physical health aspects of students, which can significantly influence their stress levels and overall well-being.
  - *headache*: Values range from 0 to 5, indicating the frequency or severity of headaches.
  - *blood_pressure*: Values are 1, 2, and 3, representing categorized levels of blood pressure.
  - *sleep_quality*: Values range from 0 to 5, reflecting the quality of sleep.
  - *breathing_problem*: Values range from 0 to 5, indicating the severity or frequency of breathing-related issues.
- **Environmental Factors**: These factors focus on the student's external environment, including aspects such as noise, safety, and living conditions, which may contribute to stress.
  - *noise_level*: Values range from 0 to 5, showing the intensity of noise in the student's environment.
  - *living_conditions*: Values range from 0 to 5, describing the quality of living conditions.
  - *safety*: Values range from 0 to 5, representing the perceived safety of the environment.
  - *basic_needs*: Values range from 0 to 5, indicating how well the student's basic needs are met.
- **Academic Factors**: These factors pertain to the student's academic experience, including performance, study load, and relationships with teachers. They help capture the stress related to academic pressures.
  - *academic_performance*: Values range from 0 to 5, reflecting the student's performance in academics.
  - *study_load*: Values range from 0 to 5, showing the amount of academic work the student is handling.
  - *teacher_student_relationship*: Values range from 0 to 5, capturing the quality of interaction between students and teachers.
  - *future_career_concerns*: Values range from 0 to 5, indicating the level of concern regarding future career prospects.
- **Social Factors**: These factors include social dynamics, such as peer pressure and support from family and friends, which can also influence the stress levels experienced by students.
  - *social_support*: Values range from 0 to 3, showing the extent of support from peers, family, and the community.
  - *peer_pressure*: Values range from 0 to 5, reflecting the degree of influence exerted by peers.
  - *extracurricular_activities*: Values range from 0 to 5, indicating the level of involvement in extracurricular activities.
  - *bullying*: Values range from 0 to 5, representing the intensity or frequency of bullying experienced.
- **stress_level** was encoded as a categorical variable (factor). "0" represents low stress level, "1" represents moderate stress level, "2" represents high stress level.

2.2 Additional Information on Scales Used

- **GAD-7 (Generalized Anxiety Disorder-7)**

  GAD-7 is a self-reported questionnaire used to screen and measure the severity of generalized anxiety disorder (GAD). It includes seven questions, with responses scored as follows: **0** for "Not at all," **1** for "Several days," **2** for "More than half the days," and **3** for "Nearly every day." The total score is calculated by summing the responses to all seven questions.

  The GAD-7 scores are interpreted as follows: a score of **0-5** indicates mild anxiety, **6-10** indicates moderate anxiety, **11-15** indicates moderately severe anxiety, and **15-21** indicates severe anxiety.

- **PHQ-9 (Patient Health Questionnaire-9)**

  PHQ–9 is a 9-qustion instrument used to screen depression and measure its severity. It takes less than 3 minutes to complete and is based on DSM-IV criteria for depression and the responses score is the same as GAD-7. The total score is calculated by summing the responses to all nine questions. The PHQ-9 scores are interpreted as follows: a score of **0-5** indicates mild depression, **6-10** indicates moderate depression, **11-15** indicates moderately severe depression, **15-21** indicates severe depression, and **21-27** indicates very severe depression.

  The PHQ-9 can be used to monitor depression severity over time and assess response to treatment, but it is not a definitive diagnostic tool.

- **The Rosenberg Self-Esteem Scale**

  The Rosenberg Self-Esteem Scale consists of 10 statements, each reflecting a different facet of self-esteem, such as self-acceptance, self-worth, and self-confidence. Half of the statements are positively worded, while the other half are negatively worded. Respondents rate each statement on a Likert scale from "strongly disagree" to "strongly agree." Points are assigned to each response, with positively worded statements scored in one manner and negatively worded statements scored in reverse. The total score, ranging from 0 to 30, indicates the individual's self-esteem level, with higher scores signifying higher self-esteem. Researchers have established thresholds to categorize respondents: 0-15 points indicate low self-esteem, 16-25 points indicate normal self-esteem, and 26-30 points indicate high self-esteem.

- **Other types of scaling**

  Many factors, including physiological, social, environmental, and academic variables such as **headache frequency**, **sleep quality**, **noise level**, **living conditions**, **academic performance**, and others, are measured on a Likert scale ranging from 0 to 5. A score of 0 generally indicates the lowest or least severe level, while a score of 5 indicates the highest or most severe level. These variables provide a standardized approach to assess subjective and environmental conditions.

  Binary variables, such as **mental_health_history**, use a straightforward 0 or 1 scoring system where 0 indicates the absence of a condition or factor, and 1 indicates its presence.

  Some factors, such as **blood pressure**, are categorized into discrete levels (e.g., 1, 2, 3) based on predefined thresholds, which may relate to clinical standards or custom criteria. Social factors such as **social support** and **peer pressure** use ordinal scales (e.g., 0–3 for social support) to measure the level of support or influence, providing an overview of social dynamics' role in stress.

2.3 Relevance to Dataset

The inclusion of variables such as **anxiety level**, **depression** and **self-esteem** in the dataset suggests that **GAD-7**, **PHQ-9** and **the Rosenberg Self-Esteem Scale** scoring systems might have been used to quantify these factors. This provides a clinically validated foundation for analyzing these psychological aspects of student stress, adding rigor and reliability to the dataset. These scales, along with their clinically validated thresholds, provide a robust framework for analyzing the dataset. By standardizing measurements, they ensure consistency and reliability in assessing the factors contributing to student stress and academic performance.

3. Objective

The primary objective of this study is to analyze and deduce insights from the dataset to better understand the factors contributing to student stress and academic performance. Using graphs and charts, we aim to uncover underlying patterns, meaningful trends,

and correlations within the data. Additionally, we will apply clustering techniques to group similar data points, gaining deeper insights into stress-related factors.

Our analysis will focus on predicting stress levels based on psychological, physiological, social, and environmental factors. Furthermore, we will extend our predictive modeling to forecast academic performance by incorporating academic factors alongside stress levels. By leveraging data visualization, clustering, and predictive modeling, this study seeks to identify key drivers of stress and academic performance, ultimately proposing proactive measures to support students effectively.

4. Data Visualization and Exploration

- The analysis of student stress factors encompasses multiple dimensions including psychological, academic, social, and physical health aspects. Through comprehensive data visualization and statistical analysis, we explore the distribution patterns, relationships, and key insights derived from the survey data. The following analysis presents a detailed examination of various stress-related factors and their interconnections among the student population.
- As **Figure 1** shows, the stress level distribution reveals a concerning pattern: The population is almost evenly split between normal, high, and very high stress levels. This indicates a significant prevalence of stress among the studied population.
- As refers to **Figure 2**, environmental factors demonstrate alarming trends, particularly in living conditions where a majority report poor conditions. Safety concerns and basic needs fulfillment show similarly troubling patterns, with nearly half reporting inadequate levels. These environmental challenges likely contribute significantly to overall stress levels.
- As refers to **Figure 3**, academic performance metrics reveal that a large portion of students are performing below average, with a particularly concerning trend in study load management. The teacher-student relationship quality shows a high percentage reporting poor relationships, though future career concerns appear more evenly distributed across the population.
- As refers to **Figure 4**, social factor analysis indicates significant challenges in support systems, with a large proportion reporting low social support. The bullying distribution shows concerning levels across all severity categories, while extracurricular participation remains notably low. These social challenges appear to create additional stress burden on students.
- As refers to **Figure 5**, physical health indicators present varied impacts, with more than half of students reporting headaches and breathing problems. Sleep quality shows a more balanced distribution across poor, average, and good categories, though the significant proportion of poor sleep quality raises concerns.

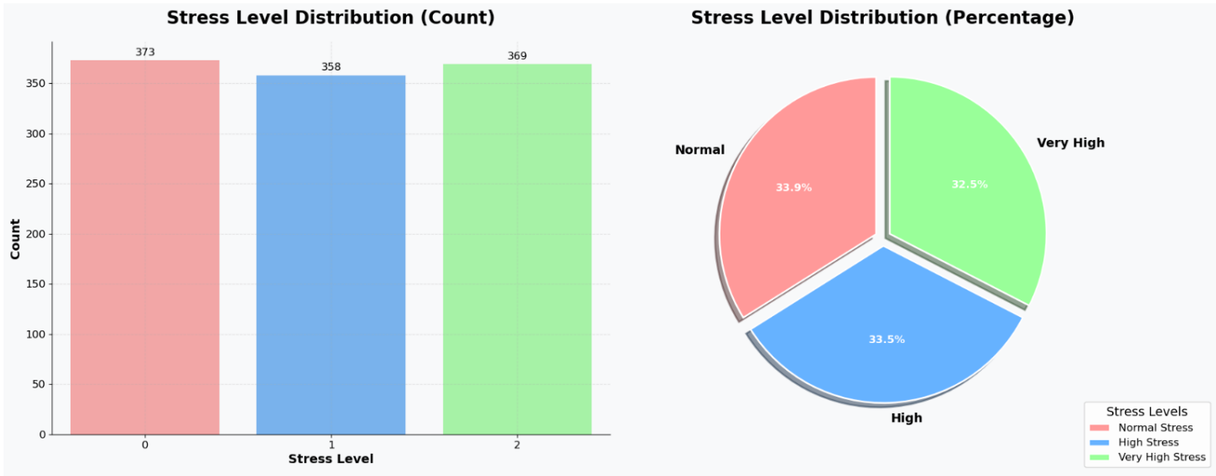

Figure 1: Distribution of stress levels across the population sample shown in count (left) and percentage (right)

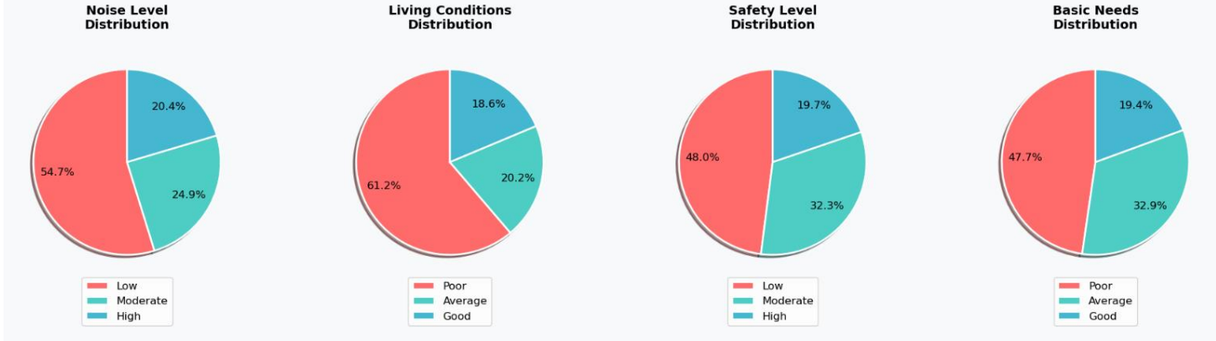

Figure 2: Distribution of environmental factors showing percentage breakdown of noise level, living conditions, safety, and basic needs

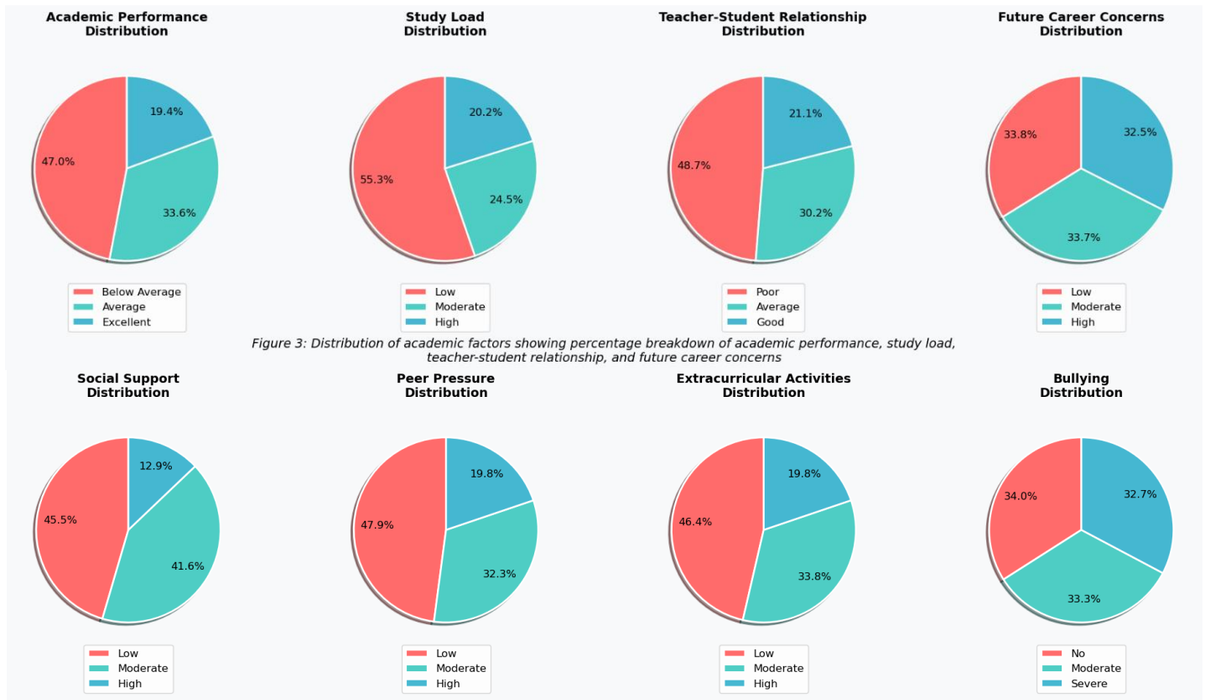

Figure 3: Distribution of academic factors showing percentage breakdown of academic performance, study load, teacher-student relationship, and future career concerns

Figure 4: Distribution of social factors showing percentage breakdown of social support, peer pressure, extracurricular activities, and bullying

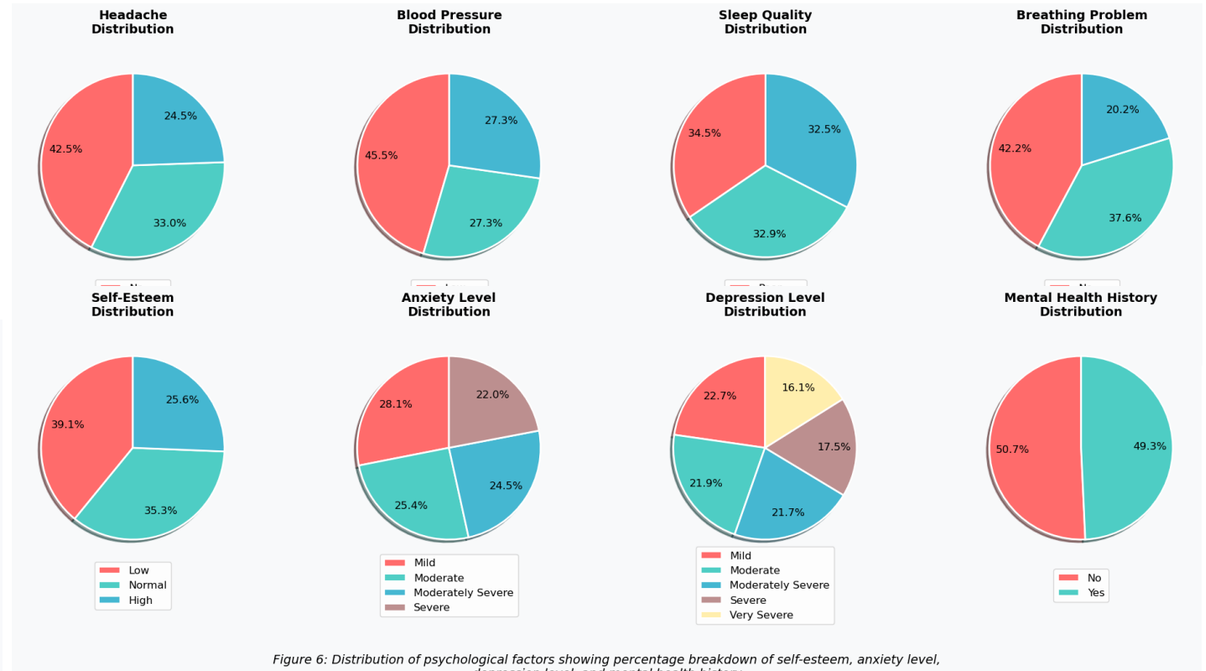

Figure 6: Distribution of psychological factors showing percentage breakdown of self-esteem, anxiety level, depression level, and mental health history

- As refers to **Figure 6**, psychological factors demonstrate concerning patterns, particularly in self-esteem levels where a substantial portion reports low ratings. The anxiety and depression distributions reveal substantial prevalence across severity levels, indicating widespread mental health challenges among the student population.
- As refers to **Figure 7**, the box plot analysis of different factors reveals similar median values across categories, though social factors show notably lower median values. This suggests a potentially systemic nature to the stress-related challenges facing students.
- As refers to **Figures 8 and 9**, the relationship between mental health history and stress levels shows strong associations, with the correlation matrix revealing strong interconnections between psychological factors and academic performance. This suggests a complex web of relationships between various stress indicators, particularly highlighting strong correlations between anxiety, depression, and overall stress levels.

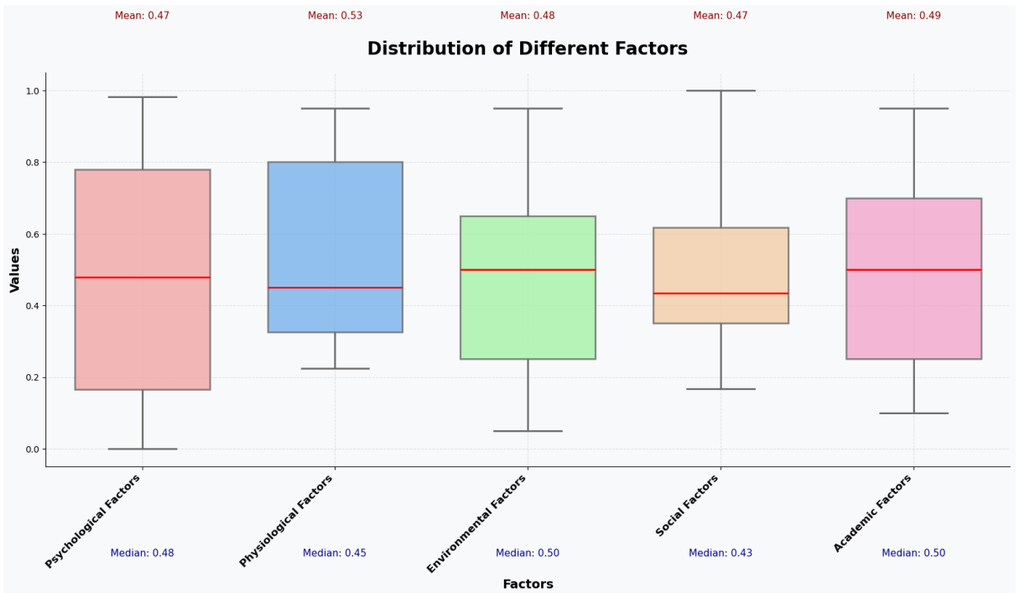

Figure 7: Box plot showing the distribution of values across different factors
Red diamonds represent outliers

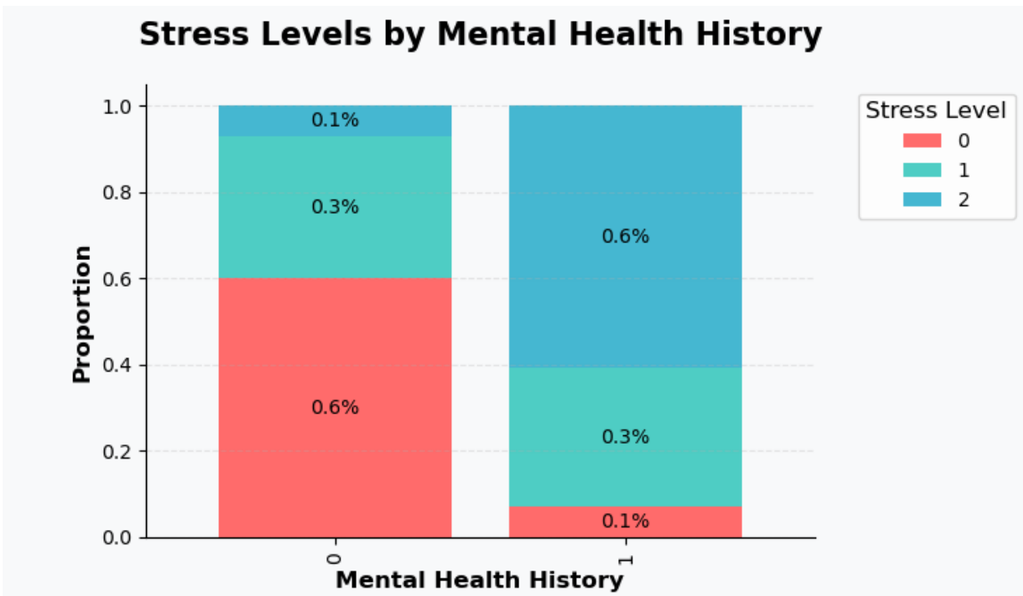

Figure 8: Distribution of stress levels across different mental health history groups

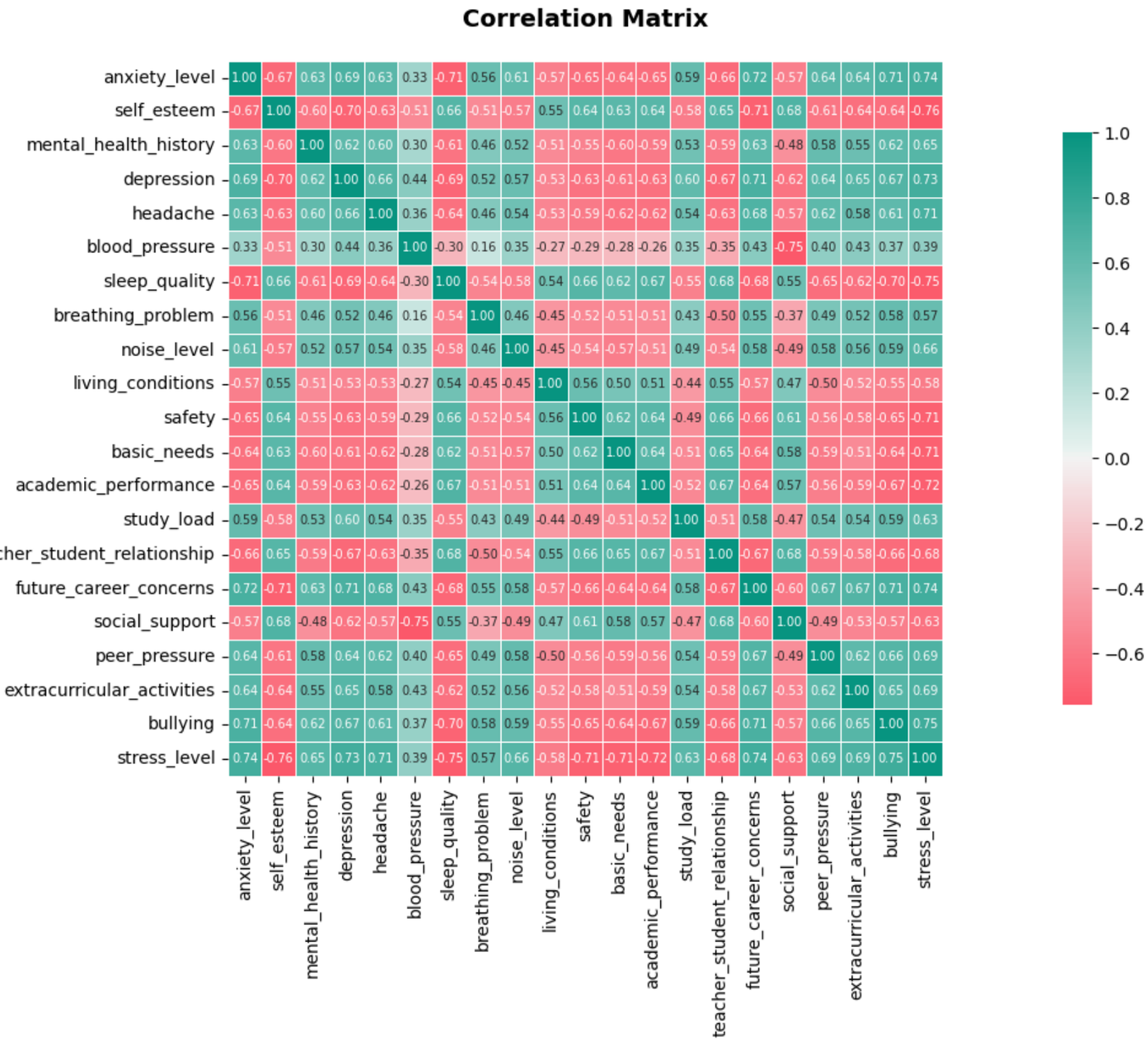

Figure 9: Correlation matrix showing the relationships between different variables in the study

5. Data Analysis

5.1 Data Preparation

The dataset was prepared by categorizing variables into psychological, physiological, environmental, academic, and social factors for focused analysis.

An initial examination of the dataset was conducted to understand its structure and statistical properties. This involved:

- **Summary Statistics**: Obtaining basic statistical measures for each variable to understand the central tendencies and variability within the data, and to check if there is any missing Variables.
- **Data Structure Inspection**: Reviewing the data types and formats of the variables to ensure they are appropriate for analysis.

As a result of these steps, No Missing Variables or duplicates were found.

To predict stress levels, a combined dataset *`psy_phy`* was created using psychological and physiological factors, which are directly related to human experience and self-reported conditions. Additionally, stress levels were predicted using social and environmental factors *`soc_env`*, which represent the surrounding conditions that are largely beyond the individual's control. Finally, a dataset combining academic factors and stress levels was used to predict academic performance, enabling an understanding of how stress interacts with academic variables. All variables in the combined datasets were appropriately formatted, such as converting categorical variables to factors where necessary. Initial correlation analysis was conducted to understand variable relationships, aiding in feature selection and guiding predictive modeling efforts. This structured approach supports targeted predictions and insights into stress and academic performance.

5.2 Student Segmentation using Clusters

Clustering is a technique used to group students based on attributes such as academic performance, stress levels, or demographic factors. This method is valuable for identifying segments of students who may exhibit similar behaviors or respond similarly to specific interventions or support strategies. By leveraging clustering, educational institutions can tailor their initiatives—such as stress management programs, academic workshops, or counseling services—to address the unique needs of each group, leading to improved outcomes and student experiences.

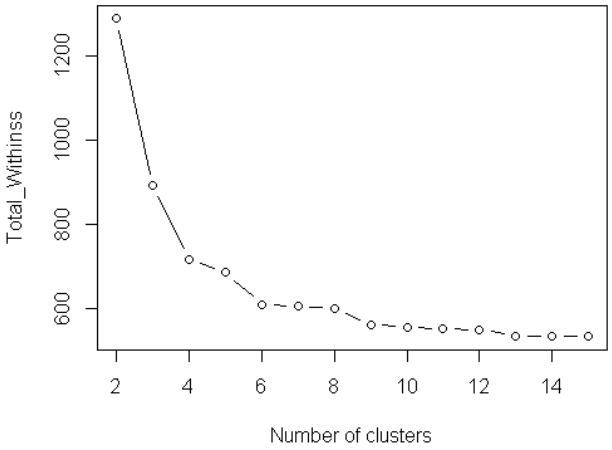

Figure 10.

Additionally, segmenting students into groups helps institutions gain a deeper understanding of their student population, enabling more informed decision-making. For example, by analyzing the characteristics and needs of each segment, targeted support can be provided to students at risk of academic difficulties or heightened stress levels, as we will explore in the analysis of our student segmentation.

Before clustering we used min-max method to normalize all the variables to have the same weights

To segment the data, we initially applied the **K-Means** Clustering technique. Different numbers of clusters were tested to determine the optimal grouping. By plotting the total within-cluster sum of squares against the number of clusters, we observed in the elbow graph in **Figure 10** that six clusters were appropriate. Increasing the number of clusters beyond six did not result in significant improvement, making six the ideal choice.

After determining the optimal cluster count, we utilized **Hierarchical Clustering** for its higher accuracy and consistent results. The analysis below provides insights into the characteristics of each cluster based on the six identified groups, as shown in **Table 1.**

Among the six identified clusters, four groups—comprising approximately 500 of the 1,100 observations—were found to be average in most factors, including stress levels. These groups differed primarily in specific factors such as social support, blood pressure, and mental health history, but overall exhibited similar levels of stress.

The remaining 600 observations were divided equally into two distinct groups of 300 students each. These groups stood out as opposites. Notably, **Cluster 2** emerged as the **most vulnerable** group, performing the worst in nearly all indicators, including anxiety,

depression, self-esteem, and academic performance. However, this group displayed the highest participation in extracurricular activities, which was an unexpected finding.

This analysis led us to the following conclusions and recommendations for this group:

- Provide mental health support, such as counseling and stress management workshops, to address their significant mental health challenges.
- Reassess their level of participation in extracurricular activities to ensure it is not contributing to burnout or exacerbating stress.
- Strengthen peer support programs and foster a more inclusive environment to reduce isolation and improve well-being.

Additionally, the data highlights a **clear relationship between engaging in more extracurricular activities and higher stress levels**, suggesting the need for a balanced approach to extracurricular commitments to promote healthier outcomes for students.

cluster no.	1	2	3	4	5	6
anxiety_level	11.34	18.05	11.47	9.75	4.19	11.16
self_esteem	19.80	7.78	20.15	15.39	27.43	14.50
mental_health_history	0.00	1.00	1.00	0.03	0.00	0.94
depression	11.83	21.38	11.46	12.58	4.10	14.36
headache	2.47	3.95	2.50	2.37	1.00	2.98
blood_pressure	1.00	3.00	1.00	3.00	2.00	3.00
sleep_quality	2.49	1.00	2.43	2.66	4.49	2.78
breathing_problem	3.00	3.95	3.05	2.51	1.51	2.30
noise_level	2.51	4.03	2.54	2.63	1.47	2.42
living_conditions	2.43	1.53	2.53	2.71	3.53	2.33
safety	2.48	1.52	2.47	2.05	4.50	2.66
basic_needs	2.55	1.53	2.48	2.57	4.52	2.21
academic_performance	2.47	1.49	2.47	2.50	4.51	2.60
study_load	2.53	3.97	2.44	2.26	1.49	2.80
teacher_student_relationship	2.46	1.49	2.51	2.03	4.47	1.77
future_career_concerns	2.45	4.51	2.47	2.58	1.00	2.64
social_support	2.50	1.00	2.55	0.56	3.00	0.56
peer_pressure	2.45	4.50	2.49	2.34	1.52	2.26
extracurricular_activities	2.44	4.47	2.53	2.83	1.49	2.21
bullying	2.58	4.46	2.55	2.37	1.00	2.36
stress_level	1.00	2.00	1.00	1.06	0.00	0.88
nom in cluster	146	300	154	110	300	90

Table (1)

5.3 Prediction Models

To achieve accurate predictions, we experimented with five different predictive models, each employing unique techniques. These models were trained and evaluated on a dataset divided randomly into training (70%) and testing (30%) subsets using the *`caTools`* library. The models were built using training data and validated on the testing data, with their accuracies computed for comparison.

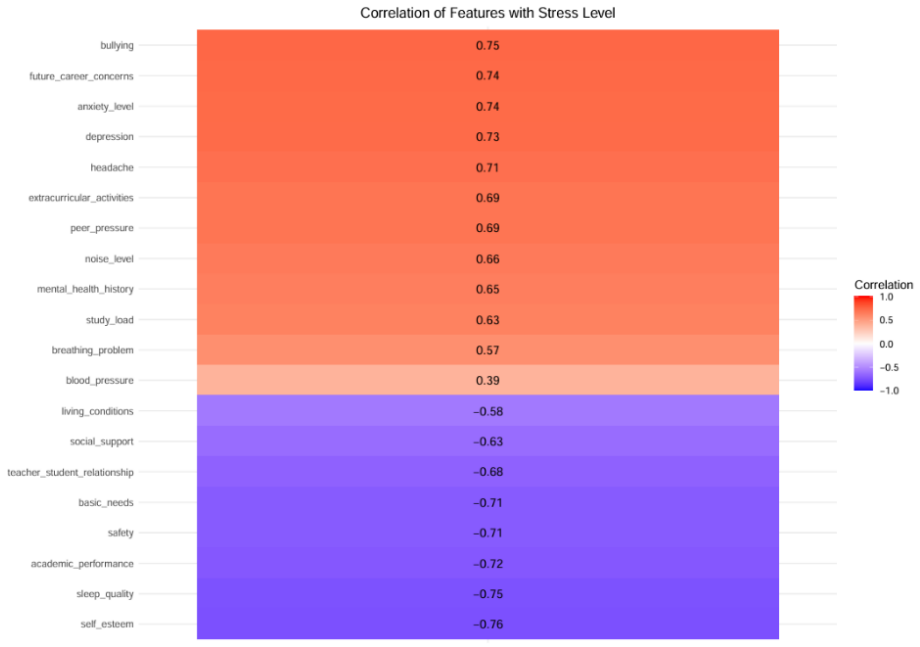

Figure 11.

Initially, predictions focused on stress levels using all variables except academic performance. To improve model performance and reduce complexity, a feature selection process was conducted based on correlation analysis:

1. **Correlation Matrix Calculation**: A correlation matrix was generated to determine the strength and direction of relationships between variables, especially their correlation with the target variable, stress_level.
2. **Highly Correlated Features**: Variables with a strong correlation (absolute correlation coefficient ≥ 0.6) with stress_level were selected as predictors. This threshold was identified as optimal after experimenting with different values (0.4, 0.5, 0.6, and 0.7), with 0.6 yielding the highest accuracy.

Our predictions after that will be based on two key categories for predicting stress:

- **Internal Factors**: Psychological and Physiological Attributes.

- **External Factors**: Social and Environmental Influences.

By integrating these factors, we aimed to analyze the primary contributors to student stress and academic performance.

Finally for predicting academic performance, it was reclassified into three categories, where "0" is below Average, "1" is Average and "2" is Excellent. This reclassification allows academic performance to be predicted accurately using other academic factors alongside the corresponding stress level.

5.3.1 Initial Machine Learning Models for Stress Prediction

5.3.1.1 CART

Classification and Regression Trees (CART) is a machine learning algorithm used for constructing decision trees. It can be applied to both classification problems (where the target variable is categorical) and regression problems (where the target variable is continuous). The CART methodology provides a recursive partitioning of the feature space to create a tree-like model of decisions that leads to the prediction of a target variable.

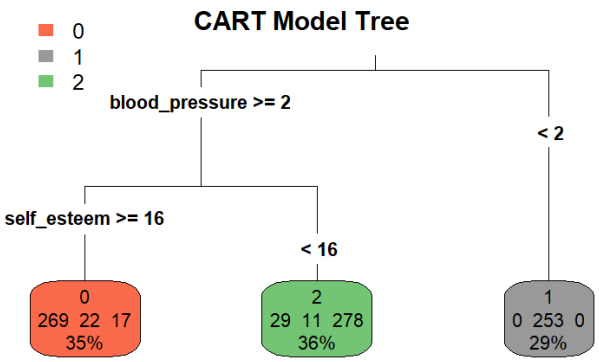

Figure 12.

An initial CART model was built using all variables (with absolute correlation coefficient ≥ 0.6) in the dataset to predict stress levels. This model identified how different variables contributed to the decision-making process. A visual representation of the tree revealed complex splits, and variable importance scores were calculated to determine the most influential factors.

Identification of Significant Variables

From the initial model, the most significant variables were identified as follows:

- Blood Pressure
- Extracurricular Activities
- Self Esteem
- Future Career Concerns
- Living Conditions
- Sleep Quality
- Social Support
- Safety

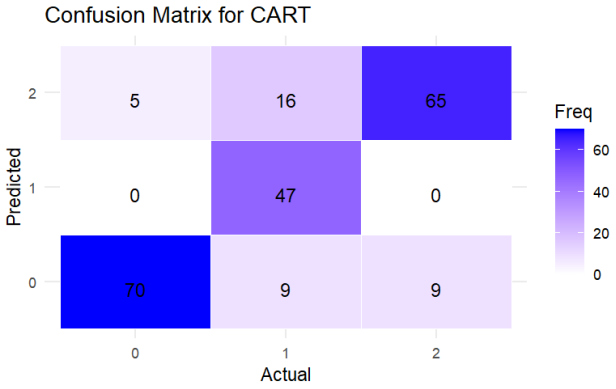

Figure 13.

These variables were used to construct a new CART model, reducing the complexity of the initial model while retaining key predictors. This model provided accuracy of **0.8235.**

Further analysis of the tree splits, as shown in Figure 12, revealed that only **blood pressure** and **self-esteem** were consistently used for decision-making. A simplified model was created using just these two variables, which also achieved an accuracy of **0.8235**. This demonstrated that the additional variables did not improve the model's performance.

This model provided accuracy of 0.8235 based on confusion matrix shown in **Figure 13**.

5.3.1.2 Random Forest

Random Forest is an ensemble machine learning algorithm that combines the predictions of multiple decision trees to produce more accurate and stable results. It is widely used for classification and regression tasks due to its high accuracy, robustness to overfitting,

and ability to handle large datasets with higher dimensionality. The performance of the Random Forest model with including all the variables done by considering variable importance as shown in Figure (14), different threshold values for variable selection, and varying the number of trees.

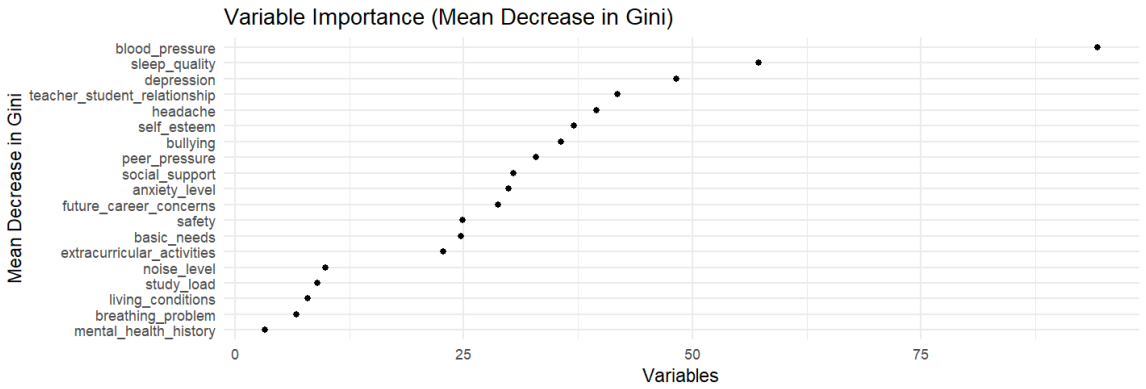

Figure 14.

1. Threshold Analysis for Variable Importance

The threshold values were applied to the MeanDecreaseGini* metric to select the most important variables. The corresponding accuracy values are as shown in the Table 2:

Threshold	Accuracy
20	0.8054
30	0.8100
35	0.8281
45	0.8235

Table (2)

Best Threshold: 35 (Accuracy: 0.8281)

Increasing the threshold beyond 35 led to a slight drop in accuracy, suggesting that including variables with lower importance can negatively impact the model's performance.

2. **Number of Trees (ntree) Analysis**

After selecting the threshold of 35, the Random Forest model was trained with varying numbers of trees (ntree) to optimize performance. The results are summarized in Table 3:

Ntree	Accuracy
800	0.8190
500	0.8281
400	0.8281
300	0.8326
200	0.8100

Table (3)

Best Number of Trees: 300 (Accuracy: 0.8326)

Increasing the number of trees beyond 300 led to overfitting, reducing accuracy.

*a metric used in Random Forest models to measure the importance of each variable in predicting the target variable. It is based on the Gini impurity, which is a criterion used to evaluate splits in decision trees.

The refined Random Forest model was built using the most important variables:

- blood_pressure
- sleep_quality
- depression
- teacher_student_relationship
- headache
- **Optimized NTREE:** 300
- The refined model's performance was evaluated using a confusion matrix shown in Figure 14.

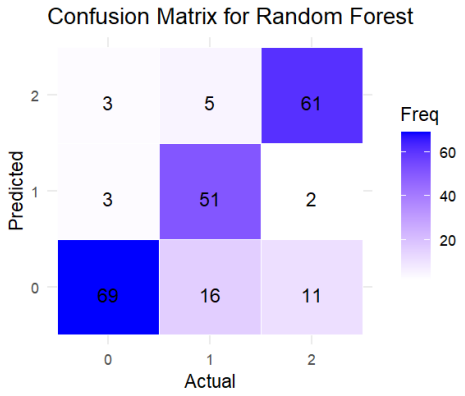

Figure 14.

5.3.1.3 Artificial Neural Networks

Neural Networks are computational models inspired by the human brain's network of neurons. In this context, a feedforward neural network was used for classification tasks, capturing nonlinear relationships between predictors and the target variable. Initially, the model was constructed with all available variables, yielding a high accuracy of **0.8281**. Subsequently, variable importance was assessed, and models were refined using only significant variables based on thresholding criteria.

Variable Importance and Threshold Analysis

Using the variable importance values derived from the model's weights, a threshold-based filtering approach was implemented. The impact of different thresholds on model performance was evaluated as shown in Table 4:

Threshold	Accuracy
0.6	0.819
0.4	0.8145
0.7	0.8054

Table (4)

The model with **Threshold = 0.6** was selected as the optimal configuration due to its simpler structure (only 8 variables) and comparable accuracy to the full model (18 variables). While the model with all variables achieved **0.8281**, the difference in accuracy was minimal, making the simpler model more practical for interpretation and deployment.

Significant variables identified:

- **self_esteem**
- **blood_pressure**
- **sleep_quality**
- **living_conditions**
- **future_career_concerns**
- **social_support**
- **peer_pressure**
- **noise_level**

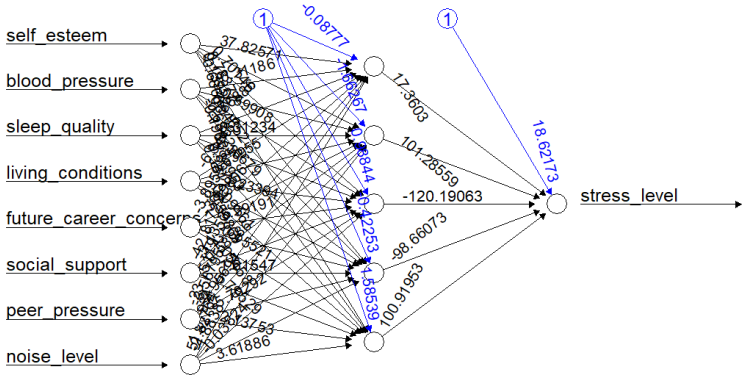

Figure 15.

The neural Network final model is visualized, presented in **Figure 15**,including its wights of the hidden layers to achieve this output.

The final model was evaluated using confusion matrix shown in **Figure 16** achieving accuracy of **0.819**

Experimenting with Hidden Layer Configurations

To optimize the model architecture, the number of hidden nodes in the network was varied (3, 5, 7, 10)

The results of this analysis were visualized in **Figure 17**, showing the relationship between hidden layer size and accuracy. The configuration with **5 hidden nodes** provided the best balance of simplicity and performance.

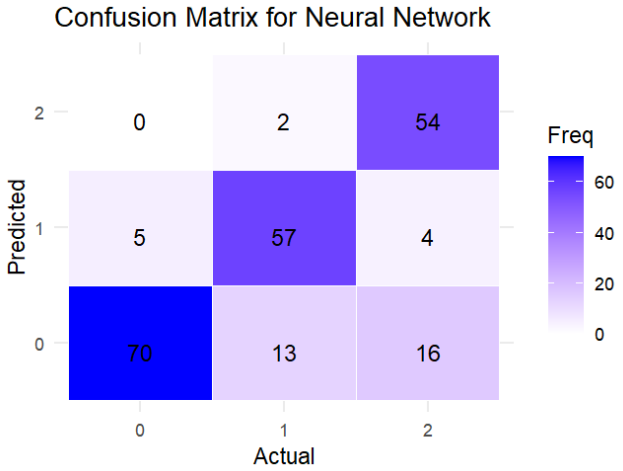

Figure 16.

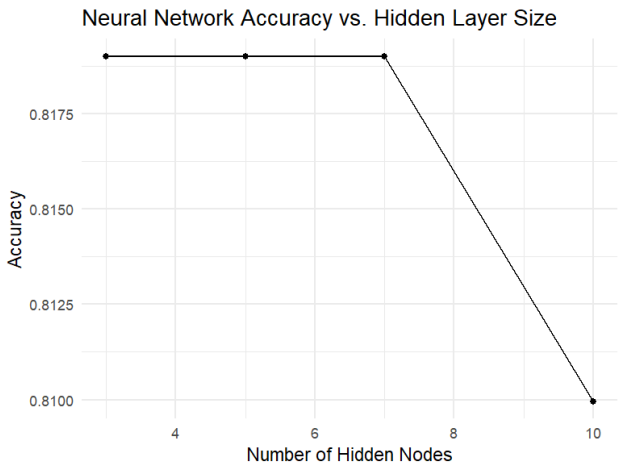

Figure 17.

5.3.1.4 Multinomial Logistic Regression

Multinomial Logistic Regression is a statistical modeling technique used when the dependent variable is categorical and consists of more than two levels (multinomial). It is an extension of binary logistic regression, allowing for the prediction of outcomes with three or more discrete categories. In the context of predicting stress levels, it can model the probability of an individual falling into different stress categories (e.g., low, medium, high) based on predictor variables.

The multinomial logistic regression model estimates the probabilities of each possible outcome of the dependent variable as a function of independent variables. It uses the **logit** link function to model the probability of each category relative to a reference category.

First, we built **Multinomial Logistic Regression** Model using all variables with **AIC = 421.9604**

Secondly, we opted to examine the correlation among the variables by systematically removing one variable at a time and observing the changes in the AIC value. We began by eliminating the variables with the highest p-values, which led to an improvement in the AIC value. Following this approach, we determined that the optimal AIC was achieved after excluding the following insignificant variables:

- Self-esteem
- Mental health history
- Headache frequency
- Blood pressure
- Noise level
- Fulfillment of basic needs
- Study load
- Quality of teacher-student relationships
- Availability of social support
- Peer pressure
- Experiences of bullying

With improved AIC = 403.645

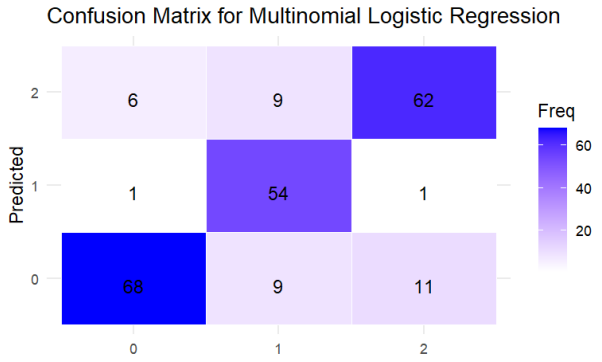

Figure 18.

The trained enhanced model was used to predict stress levels on the testing set. A confusion matrix was generated, shown in Figure 18, to assess the model's performance, with an Accuracy = 0.8326

5.3.1.5 Support Vector Machine

Support Vector Machine (SVM) is a powerful machine learning algorithm designed for both classification and regression tasks. It is particularly well-suited for solving non-linear problems by mapping data into a higher-dimensional space using kernel functions. The primary goal of SVM is to find the optimal hyperplane that separates data points into different classes with the maximum margin, ensuring robust generalization to unseen data.

In the context of predicting stress levels, the SVM model was applied to classify individuals into distinct stress categories (e.g., low, medium, high). This technique was chosen due to its ability to handle high-dimensional spaces effectively and its resilience to overfitting, especially in cases with fewer data points relative to the number of features.

The trained model was used to predict stress levels on the testing set. A confusion matrix was generated, shown in Figure 19, to assess the model's performance, with an Accuracy = 0.8326

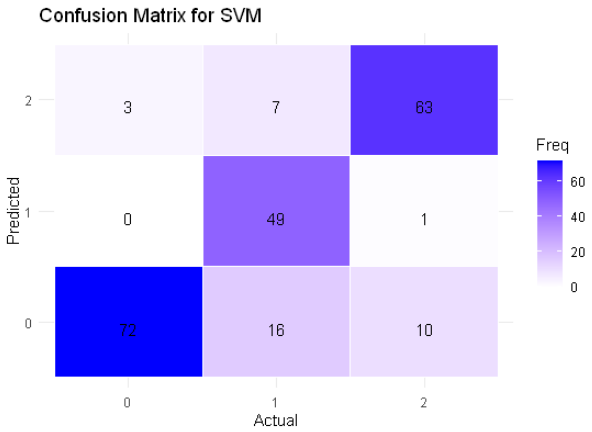

Figure 19.

5.3.1.6 Initial Results

The performance of different models for predicting stress levels is summarized in the Table 5:

Model	Accuracy
CART	0.8235
Random Forst	0.8326
Neural Networks	0.819
Multinominal Logistic Regression	0.8326
Support Vector Machine	0.8326

Table (5)

The performance of the models for predicting stress levels shows that Random Forest, Multinomial Logistic Regression, and Support Vector Machine all achieve the highest accuracy, making them the top-performing models in this comparison. While the CART model follows with a slightly lower accuracy, it may still be valuable for its simplicity and interpretability.

Overall, Multinomial Logistic Regression might be the best option as it's the least complex model and has highest prediction accuracy.

5.3.2 Predicting Stress Levels using Internal Factors

Using the same models as before, we predicted outcomes based solely on psychological and physiological factors, excluding their direct correlation with the target variable. This approach aimed to assess how well internal factors alone could contribute to the model's performance.

To ensure the reliability of the results and account for the randomness of dataset splitting, we validated each of the five models using over **1,000 different random seeds.** The aggregated results are summarized in Table 6. The results indicate that Multinomial Logistic Regression achieved the highest average performance with relatively low variability, making it the most consistent and reliable model for predicting outcomes based solely on internal factors. While Random Forest also demonstrated strong performance (average = 0.8866) and stability, the Neural Network showed the least average and the highest variability, suggesting sensitivity to dataset splits and potential overfitting issues. Overall, models leveraging simpler architectures, such as Multinomial Logistic Regression and Random Forest, are better suited for predictions in this context due to their robust performance and lower sensitivity to randomness.

Model	Average	Std. Dev	Max	Min
CART	0.8813	0.0143	0.9273	0.8303
Random Forest	0.8866	0.0151	0.9364	0.8333
Neural Network	0.8749	0.0432	0.9364	0.3273
Multinomial Logistic Regression	0.8910	0.0148	0.9333	0.8455
Support Vector Machine	0.8835	0.0150	0.9273	0.8364

Table (6)

5.3.3 Predicting Stress Levels using External Factors

We analyzed the predictive performance of the same models by focusing on a different set of predictors. As before, the objective was to assess how well these factors contribute to outcomes when used independently in the models. To ensure the reliability of results, we validated each model across 1,000 different random seeds, reducing the impact of randomness in dataset splitting.

The results are summarized in Table (7).

The results indicate that Random Forest achieved the highest average performance and the lowest variability, establishing it as the most consistent and reliable model for predicting outcomes based solely on external factors. Multinomial Logistic Regression also demonstrated strong performance (average = 0.8782) and stability, making it a competitive alternative. In contrast, the Neural Network once again showed the lowest average performance and the highest variability, highlighting its sensitivity to dataset splits and potential overfitting issues.

Model	Average	Std_Dev	Max	Min
CART	0.8774	0.0147	0.9182	0.8152
Random Forest	0.8797	0.0144	0.9273	0.8394
Neural Network	0.8763	0.0163	0.9242	0.8091
Multinomial Logistic Regression	0.8782	0.0149	0.9182	0.8242
Support Vector Machine	0.8777	0.0150	0.9242	0.8273

Table (7)

Overall, Random Forest stands out as the most suitable model in this context, combining robust performance with low sensitivity to randomness. Notably, its predictive accuracy can reach as high as 92.73%, underscoring its effectiveness in leveraging external factors for accurate predictions.

5.3.4 Comparative Insights

- **Stronger Performance with Internal Factors**: Models generally performed better when using internal factors.
- **Consistency of Random Forest**: Random Forest consistently delivered strong and stable performance across different datasets, making it a versatile and reliable model.
- **Sensitivity of Neural Networks**: The Neural Network's variability was evident in both cases, requiring careful tuning and preprocessing for optimal performance.

- **Robustness of Multinominal Logistic Regression and CART**: These simpler models remained reliable across both scenarios, balancing good performance and low variability.

5.3.5 Predicting Academic Performance

Academic performance prediction is a crucial aspect of understanding student success and identifying factors that can impact their learning outcomes. The goal was to use various predictive models to evaluate how well academic and non-academic factors could classify students into these categories, providing insights into key drivers of academic success and areas for targeted intervention.

As with the stress level models, the dataset was randomly split into 70% for training and 30% for testing, with the Academic Performance variable as the target. And all variable data types were converted from integer to factor to better suit the modeling process.

5.3.5.1 Initial Naive Approach

The goal is to predict students' academic performance using all various factors (with absolute correlation coefficient ≥ 0.6). This will involve applying both CART and Random Forest models to assess the initial accuracy of the predictions. By focusing on factors with strong correlations, this approach aims to provide an early indication of how well these models can predict academic performance, establishing a baseline for further refinement and optimization.

The results from the initial naive approach were quite poor, with the CART model achieving an accuracy of only 43.5%, and the Random Forest model performing slightly better at around 48%. These low accuracies indicate that, despite using factors with strong correlations to academic performance, the models did not capture the underlying relationships effectively. This suggests that additional steps, such as refining the feature selection, tuning the models, or incorporating more relevant factors, may be necessary to improve the predictive performance.

5.3.5.2 Reclassified Approach for Improved Accuracy

In the next phase, the academic performance variable was reclassified into three categories to simplify the prediction task. The original six levels (0-5) were grouped as follows: performance levels 0 and 1 were combined into "0" (Below Average), levels 2 and 3 into "1" (Average), and levels 4 and 5 into "2" (Excellent). This reclassification aimed to reduce complexity and provide a more straightforward categorization, potentially improving the models' accuracy in predicting academic performance. By focusing on broader performance categories, the models could more easily identify patterns and make more reliable predictions. While this approach simplified the problem, it came at the cost of losing some fine-grained information about the students' exact performance levels. However, the hope was that by consolidating the categories, the models would achieve higher levels of accuracy, as the broader categories might help the algorithms focus on more prominent patterns without being overly sensitive to small differences in performance. The aim of this analysis was not to predict students' exact academic levels, but rather to understand how factors such as stress and academic-related variables influence their performance. Additionally, unlike the naive approach, this model will be based only on stress and academic factors, regardless of their correlation, rather than including all available variables.

The same five models used in stress prediction were applied to predict academic performance based on stress and academic-related factors. These models included CART, Random Forest, Neural Network, Multinomial Logistic Regression, and Support Vector Machine, all of which were evaluated to assess their effectiveness using over 1,000 different random seeds. The results were as shown in Table 8:

Model	Average	Std_Dev	Max	Min
CART	0.7300	0.0148	0.7697	0.6788
Random Forest	0.7306	0.0181	0.7788	0.6697
Neural Network	0.7369	0.0189	0.7939	0.6697
Multinomial Logistic Regression	0.7261	0.0186	0.7788	0.6576
Support Vector Machine	0.7291	0.0165	0.7697	0.6758

Table (8)

The results suggest that while all five models performed reasonably well in predicting academic performance based on stress and academic factors, the **Neural Network** and **Random Forest** models were the most effective, with **Neural Network** slightly outperforming the others in terms of accuracy. These models were able to capture the underlying relationships in the data and provide more reliable predictions, which may be crucial for understanding how stress and academic factors influence student performance. However, the differences in performance across models also highlight the importance of model selection and tuning to optimize predictions for specific tasks.

6. References

1. "GAD-7 and PHQ-9: Clinical Tools." *Medi-Stats*, www.medi-stats.com/gad-7-phq-9.
2. Kroenke, Kurt, et al. "The PHQ-9: Validity of a Brief Depression Severity Measure." *Journal of General Internal Medicine*, vol. 16, no. 9, 2001, pp. 606–613, doi:10.1046/j.1525-1497.2001.016009606.x.
3. Rosenberg, Morris. "The Rosenberg Self-Esteem Scale." Verywell Mind, https://www.verywellmind.com/the-rosenberg-self-esteem-scale-8699962.
4. Iowa State University Extension and Outreach. *Likert Scale Examples for Surveys*. Iowa State University, n.d. Web. 28 Dec. 2024. https://bit.ly/3Q1kn67
5. UCLA Statistical Consulting Group. "Multinomial Logistic Regression." *UCLA Institute for Digital Research and Education*, 2021, https://stats.oarc.ucla.edu/r/dae/multinomial-logistic-regression/. Accessed 28 Dec. 2024.
6. GeeksforGeeks. "Support Vector Machine Algorithm." *GeeksforGeeks*, 2021, https://www.geeksforgeeks.org/support-vector-machine-algorithm/. Accessed 28 Dec. 2024.