Student Stress Factor Detection Through Classification and Feature Selection

Anubhav Mohanty

Regd. No. 2001209087

8th, CSE

anubhavrs@gmail.com

Sagar Verma

Regd. No. 2001209168

8th, CSE

cse.20bse64@silicon.ac.in

Rajat Kumar

Regd. No. 2001209058

8th, CSE

rajatkr.0225@gmail.com

Payal Pani

Regd. No. 2001209147

8th,CSE   
panipayal232002@gmail.com

Supervised by: Prof.(Dr) Pamela Chaudhury

***Abstract* — *This project focuses on addressing student stress, a significant concern impacting academic performance and well-being, through advanced machine learning techniques. By leveraging classification, feature selection, and optimization methods, the goal is to develop a robust model capable of accurately identifying stress factors across various domains. To enhance the dataset's diversity and relevance, additional samples from Kaggle are incorporated. The project emphasizes the importance of classification and feature optimization in creating a comprehensive stress detection model, with evaluation based on meticulous analysis of performance metrics to ensure accuracy and reliability. An interactive interface is introduced to overcome limitations in existing works, providing students with a personalized stress detection experience and a feedback mechanism. Overall, the project aims to improve real-world applicability by designing the dataset to reflect the present scenario accurately.***

***Keywords — Student Stress Detection, Classification Algorithms***

# Introduction To the Project

In recent years, the recognition of stress's significant impact on students' academic performance and well-being has grown. From academic pressures to social expectations and personal challenges, educational settings are rife with stressors, demanding effective strategies for identification and mitigation.

This project aims to meet this challenge by developing a sophisticated system for student stress detection, employing advanced machine learning techniques. By analyzing diverse data, including academic records and demographic information, the project seeks to offer educators invaluable insights for targeted interventions.

Traditional methods of assessing student stress often lack scalability and precision, emphasizing the necessity of a data-driven approach. Through the utilization of machine learning and optimization techniques, the project endeavors to enhance stress detection accuracy, facilitating more personalized and efficient interventions.

The urgency of this endeavor is heightened by the recent surge in student stress, exacerbated by the pandemic. By surmounting existing limitations and creating an interactive interface, the project endeavors to craft a comprehensive stress detection model that empowers students to manage stress levels and elevate overall well-being.

# Dataset Details

Securing a comprehensive and dependable dataset that aligns with our project goals presented a considerable challenge. Many of the initial datasets we encountered either lacked essential features or were limited in sample size. Fortunately, we came across the dataset titled "Student Stress Factors: A Comprehensive Analysis" on Kaggle. This dataset encompasses survey responses collected from 1,100 students aged between 15 to 24. The wide age range covered both high school and college students, ensuring inclusivity. Within this dataset, a total of 21 features were provided, with 20 contributing to each student's overall stress factor. These features were categorized into five major groups and were measured using specific numerical ranges or popular metrics. The categories encompassed psychological, physiological, environmental, academic, and social factors.

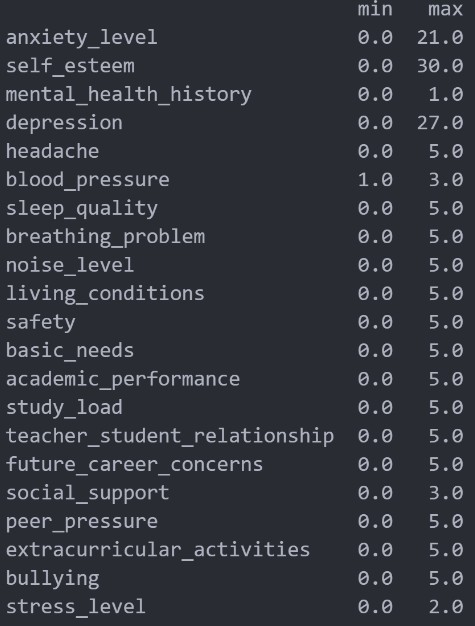


Fig. 1. Stress Factor Dataset Features

## Psychological Factors

This category comprises various factors such as anxiety level, self-esteem, mental health history, and depression. The anxiety level of a student is assessed using the GAD-7 score. This scoring system involves presenting seven questions to students, each requiring a numerical response ranging from 0 to 3. The responses are then summed up, yielding a final score between 0 and 21. The classifications based on these scores are as follows:

1) 0-4: minimal anxiety

2) 5-9: mild anxiety

3) 10-14: moderate anxiety

4) 15-21: severe anxiety

Self-esteem is evaluated using the Rosenberg Self-Esteem Scale. This scale utilizes a complex scoring system that determines an individual’s self-esteem based on their responses to a series of questions. Each question elicits a response ranging from 4 to 1, indicating strongly agree, agree, disagree, and strongly disagree, respectively. Points corresponding to the response and question number are tallied up to a maximum of 30, with higher scores indicating higher self-esteem. Mental health history in this dataset is binary, with a value of 0 indicating no history of mental health issues and 1 indicating a history. Depression is measured using the Patient Health Questionnaire, or PHQ-9. This assessment comprises several questions, each answered numerically from 0 to 3. The final score, ranging up to 27, is derived by summing up the numerical answers. The classifications based on these scores are:

1) 1-4: Minimal depression

2) 5-9: Mild depression

3) 10-14: Moderate depression

4) 15-19: Moderately severe depression

5) 20-27: Severe depression

## Physiological Factors

In the physiological category, factors such as headache, blood pressure, sleep quality, and breathing problems are considered. Unlike the psychological features, the scaling for these factors varies significantly. Headache, sleep quality, and breathing problems are rated on a scale from 0 to 5 , with the following interpretations:

1) 1: Low

2) 2: Medium-low

3) 3: Medium

4) 4: Medium-high

5) 5: High

Blood pressure, on the other hand, was assessed on a scale from 1 to 3. Initially confusing, it was later understood that the participants' actual blood pressure was classified into one of three categories:

* + 1. 1: Low
    2. 2: Normal
    3. 3: High.
  1. *Environmental Factors*

In the environmental category, various environmental factors that students commonly encounter are considered. These include noise level, living conditions, safety, and basic needs. All four features are assessed on a scale from 0 to 5, with the same classification system previously mentioned.

Noise level indicates the level of loudness or distractions in the student's environment. Living conditions evaluate the quality of the student's living situation. Safety assesses the sense of safety the student feels both in their living environment and in general. Lastly, basic needs gauge whether the student's basic living needs are being met, such as having an adequate number of meals per day.

## Academic Factors

In the academic category, various factors within the student's academic environment that could impact their stress levels are considered. These factors include academic performance, study load, teacher-student relationship, and future career concerns. All four features are assessed on the same scale from 0 to 5, as previously mentioned.

Academic performance provides a general measure of the student's performance and their grades in school. Study load evaluates the frequency of studying and the overall workload of the student's classes. The teacher-student relationship assesses the quality of the academic relationship between the student and their teacher, considering whether it is positive or negative. Initially appearing unusual, this feature is justified by the significant impact a poor relationship with a teacher can have on a student's stress levels. Lastly, future career concerns measure the level of worry the student has about their prospects after completing their education. While this feature may seem unconventional at first, it is reasonable considering how concerns about post-graduation life can contribute to stress levels.

## Social Factors

The social category focuses on assessing how social factors, both within and outside of the school environment, can impact a student's stress level. This category includes social support, peer pressure, extracurricular activities, and bullying. These features are evaluated on a scale from 0 to 5, except for social support, which is rated from 0 to 3.

Social support measures the level of support the student receives from friends or family, encompassing practical assistance or emotional support. Peer pressure evaluates the extent to which students are influenced by their peers to engage in certain behaviors or adopt particular viewpoints. Extracurricular activities gauge the student's level of involvement in activities outside of their academic curriculum, such as clubs, sports, or hobbies. Lastly, bullying assesses the frequency of the student's experiences with bullying, whether within or outside of the school environment.

We made the deliberate decision to utilize all 20 features contributing to the overall stress level, as each of these factors could potentially play a crucial role in determining a student's stress levels. We chose not to exclude any features initially, as we believed that even those seemingly less significant might prove relevant in stress level prediction. Our approach assumed that any unimportant features would have negligible impact on stress level prediction and could be disregarded.

Upon discovering and examining the dataset, we were pleasantly surprised to find that minimal data cleaning was necessary. All 1,100 entries were complete, with responses recorded for each feature. This lack of missing data or need for row removal was a welcome discovery. However, we encountered one initial challenge regarding the scaling of each feature. Unfortunately, the dataset lacked thorough documentation, leaving us uncertain about the interpretation of the numerical values. Fortunately, after several weeks, the dataset's author responded to a query on Kaggle, providing clarification on the measurements and updating the documentation. This timely assistance was instrumental, as it enabled us to proceed with confidence in utilizing the dataset.

Although data cleaning was minimal, we noticed a slight imbalance in the distribution of stress level classifications even after adding 308 additional data points to the dataset bring it to 1408 data points. Specifically, there were 411, 540, and 457 students categorized into stress levels 0, 1, and 2, respectively.

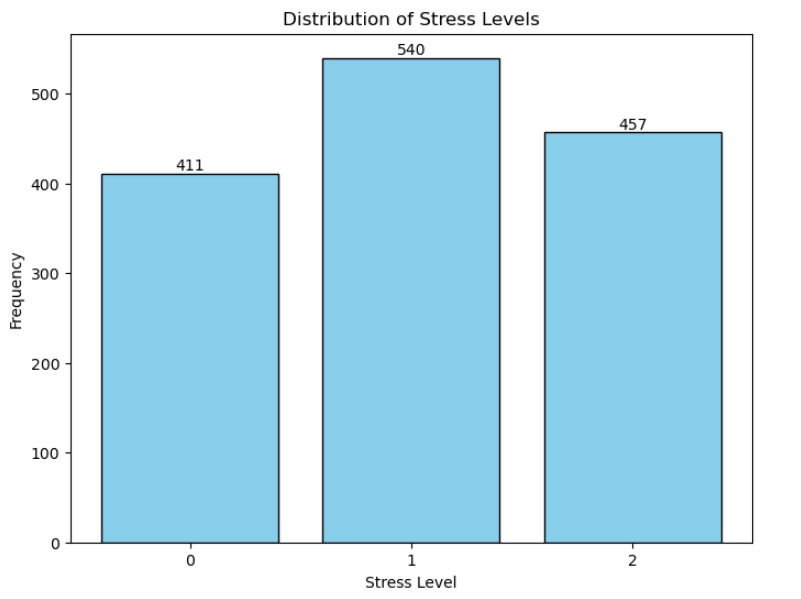


Fig. 2. Bar Chart: Number of Students in Each Stress Level Category

Although we don't anticipate the slight imbalance in stress level classifications to significantly impact our model's results, we're curious to explore its potential effects on performance. Despite this imbalance, we remain confident in the dataset's suitability for our project.

The dataset's abundance of features made it ideal for our purposes. Once the documentation was clarified and the feature measurement scales became clearer, we found the dataset to be intuitive and well-suited for our analysis. In conclusion, we are satisfied that we selected the ideal dataset for our project.

# Approach and Methodology for Machine Learning

For our project, we opted for five machine learning methodologies to address both research questions. These methodologies encompassed the utilization of K-Nearest Neighbor, Logistic Regression, Decision Tree, Gaussian Naive Bayes and Artificial Neural Network. In each case, we conducted a training-testing split of 80% and 20%, respectively, maintaining a consistent random state of 42 to ensure reproducibility. By employing these six approaches, we aimed not only to classify new students into one of three stress level categories based on their factors but also to determine which features exert the most significant influence on a student's stress level.

## K-Nearest Neighbors

#### To start addressing our research question about predicting a new student's stress level based on their stress factors similarity to known students, we began with K-Nearest Neighbors (KNN). This method stores the training dataset and, when given the testing data, identifies the closest training example to classify the testing instance into the same stress level category. These categories are represented by integers 0, 1, or 2, denoting different stress levels. We chose KNN initially due to its simplicity and straightforward hyperparameter tuning process.

#### We evaluated the KNN model's performance by assessing its accuracy in correctly classifying the testing data into their respective stress level categories, as determined by the training data. The model's accuracy was determined using the "accuracy\_score" metric from sklearn's metric package. The results obtained were plotted for future comparisons.

#### To optimize our model's performance, we employed K Fold-Cross Validation with ten splits for hyperparameter tuning. This involved testing odd numbers ranging from 1 to 30 as the number of neighbors. Each iteration involved partitioning the training data into smaller subsets for validation. We then ran the KNN classifier with the current iteration's number of neighbors, predicted the validation set, and calculated the accuracy score. Through this iterative process, we determined that 27 neighbors yielded the optimal performance.

## Decision Tree

To explore our research question concerning the significant factors contributing to predicting student stress levels, we selected Decision Trees for their interpretability. This machine learning technique employs a hierarchical tree-like structure to systematically evaluate and select features based on their importance in predicting the target variable—in this case, student stress levels. We configured the Decision Tree classifier with no maximum depth to fully expand the tree and implemented the entropy criterion to measure information gain. Additionally, we fixed the classifier's random state at 42 for reproducibility.

Following training, we assessed the model's performance on unseen test data using accuracy metrics. The accuracy of the model was determined by calculating the "accuracy\_score" provided by sklearn's metric package. Furthermore, we extracted feature importance from the Decision Tree's "feature\_importances\_" property to gain deeper insights into the influential variables.

Recognizing the importance of fine-tuning our model for optimal performance, we conducted hyperparameter tuning, focusing on the "max\_depth" parameter of the Decision Tree classifier. This iterative process involved testing various values to identify the one that maximizes accuracy while mitigating overfitting. Subsequently, we re-ran the classifier with the refined parameter settings and documented both accuracy and feature importance.

## Logistic Regression

In pursuit of understanding the significant factors influencing student stress levels, we turned to Logistic Regression for its interpretability and simplicity. Logistic Regression is a linear model that analyzes the relationship between the independent variables (stress factors) and the dependent variable (stress levels) by estimating probabilities using a logistic function. We configured the Logistic Regression model to optimize interpretability by not imposing regularization and ensuring convergence with a high tolerance value. Additionally, we set a fixed random state of 42 to maintain reproducibility throughout our analysis.

After training the model, we evaluated its performance on unseen test data using accuracy metrics. The accuracy of the Logistic Regression model was assessed by computing the "accuracy\_score" provided by sklearn's metric package. Furthermore, we delved into the feature coefficients to discern the importance of each stress factor in predicting student stress levels.

Acknowledging the necessity of fine-tuning our model for optimal performance, we engaged in hyperparameter tuning, focusing on parameters such as regularization strength and solver algorithm. This iterative process involved experimenting with different parameter values to identify the optimal configuration that maximizes accuracy while preventing overfitting. Subsequently, we retrained the Logistic Regression model with the refined parameter settings and documented both accuracy metrics and feature coefficients for further analysis.

* 1. *Gaussian Naïve Bayes*

In our quest to uncover the critical factors shaping student stress levels, we turned to Gaussian Naïve Bayes (GNB) for its simplicity and intuitive probabilistic framework. GNB is a probabilistic classifier based on Bayes' theorem with the assumption of independence between features. Despite its simplifying assumption, GNB often performs well in practice, especially with continuous-valued features.

We configured the GNB model to handle continuous features by assuming a Gaussian distribution for each class. This allowed us to model the probability density function of each feature independently for each class, enabling efficient classification.

After training the GNB model, we evaluated its performance on unseen test data using standard metrics such as accuracy. The accuracy of the GNB model was computed using the "accuracy\_score" function provided by sklearn's metric package.

Furthermore, we explored the importance of each feature in predicting student stress levels by analyzing the class conditional probabilities estimated by the GNB model. This provided insights into which features contribute most significantly to the classification decision.

Recognizing the importance of fine-tuning our model for optimal performance, we engaged in hyperparameter tuning, focusing on parameters such as priors and variances. This iterative process involved experimenting with different parameter values to identify the optimal configuration that maximizes accuracy.

Subsequently, we retrained the GNB model with the refined parameter settings and documented both accuracy metrics and feature importance for comprehensive analysis and interpretation.

* 1. *Artificial Neural Network*

In our endeavor to unveil the pivotal factors influencing student stress levels, we delved into Artificial Neural Networks (ANNs) for their ability to capture complex relationships within data. ANNs are a class of machine learning models inspired by the biological neural networks of the human brain. These models consist of interconnected nodes (neurons) organized in layers, allowing them to learn intricate patterns and representations from data.

We configured the ANN model with multiple hidden layers to enable it to learn and represent hierarchical features present in the stress factors. Each neuron within the network applies a weighted sum of inputs followed by a nonlinear activation function, enabling the model to capture nonlinear relationships between features and stress levels.

After training the ANN model, we evaluated its performance on unseen test data using standard metrics such as accuracy. The accuracy of the ANN model was computed using the appropriate evaluation metric provided by sklearn's metric package.

Furthermore, we conducted feature importance analysis by examining the learned weights and biases within the network. This provided insights into which features contribute most significantly to predicting student stress levels.

Recognizing the importance of optimizing our model for optimal performance, we engaged in hyperparameter tuning, focusing on parameters such as the number of hidden layers, number of neurons per layer, and choice of activation functions. This iterative process involved experimenting with different configurations to identify the optimal architecture that maximizes accuracy while avoiding overfitting.

Subsequently, we retrained the ANN model with the refined parameter settings and documented both accuracy metrics and feature importance for comprehensive analysis and interpretation.

# Results

In the Results section, I present the outcomes derived from my investigation into predicting and comprehending student stress levels employing a diverse ensemble of machine learning models. Addressing the research questions, "Can we predict the stress level of a new student based on the similarity of their stress factors to those of known students?" and "Which factors contribute most significantly to predicting stress levels in students?", I utilized six distinct models: K-Nearest Neighbor, Logistic Regression, Decision Tree, Gaussian Naive Bayes and Artificial Neural Network. Each model was chosen based on its specific strengths and capabilities to address different aspects of the research inquiries effectively.

These models were carefully selected to leverage their unique features and functionalities in extracting meaningful insights from the dataset. The subsequent analysis presents a comprehensive overview of their performance and contributions towards understanding student stress dynamics, highlighting the importance of employing diverse modeling approaches to attain comprehensive insights.

## K-Nearest Neighbors

KNN (K-Nearest Neighbors) classification was utilized in this research paper because of its simplicity and flexibility. KNN is a non-parametric algorithm that does not make any assumptions about the underlying data distribution and can handle both continuous and categorical data.

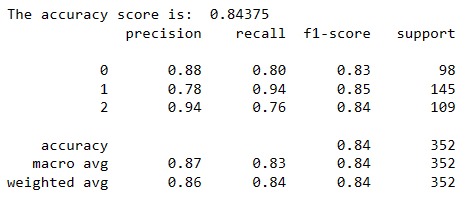


Figure 3. Classification report of KNN Classification

## Decision Tree

Decision Tree Classification models the relationship between independent and dependent variables by recursively partitioning data into subsets, aiming to maximize impurity reduction. It is versatile, handling both binary and multi-class classification and suitable for non-linear relationships and interactions between variables. Additionally, it can manage missing data and outliers effectively.

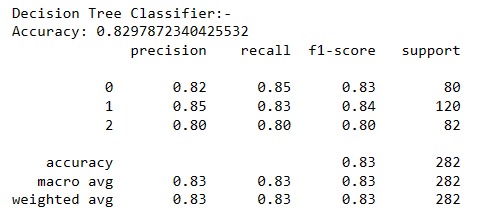


Figure 4. Classification report of Decision Tree Classification

## Logistic Regression

Logistic regression is a popular and widely used algorithm for binary classification problems, where the goal is to predict one of two possible outcomes. In this research paper, logistic regression was chosen for the binary classification task due to its simplicity, interpretability, and effectiveness in handling linearly separable datasets. Logistic regression estimates the probability of the positive class as a function of the input variables, using a sigmoid function that outputs values between 0 and 1 predicting whether or not the person has Parkinson.

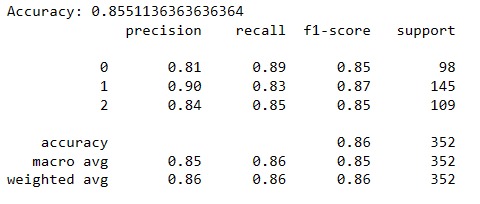


Figure 5. Classification report of Logistic Regression

* 1. *Gaussian Naïve Bayes*

Naive Bayes classification was used due to its simplicity and effectiveness in handling high-dimensional datasets. Naive Bayes is a probabilistic algorithm that is based on the Bayes theorem and the assumption of conditional independence between features. This assumption simplifies the probability calculation by allowing the algorithm to estimate the probability of each feature independently, given the class variable.

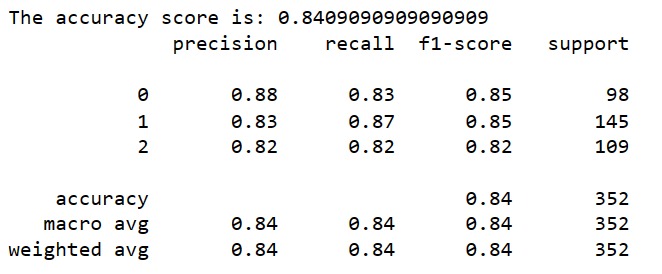


Figure 6. Classification report of Naïve Bayes Classifier

* 1. *Artificial Neural Network*

The neural network architecture is defined with an input layer, two hidden layers employing ReLU activation functions, and an output layer utilizing softmax activation for multiclass classification. Following model compilation with the Adam optimizer and categorical cross-entropy loss function, training commences on the training data. Once trained, the model predicts probabilities for the test data. True labels are binarized for evaluation, and accuracy is computed by comparing predicted labels to true labels. Ultimately, the code prints the accuracy score, providing a measure of the model's performance in classifying unseen data.

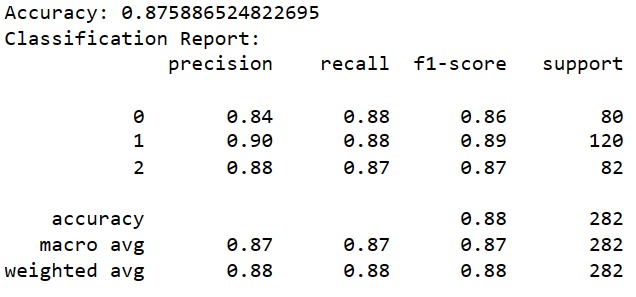


Figure 7. Classification report of Neural Network

# Conclusion

The conclusion of this project highlights the comprehensive evaluation of key factors contributing to student stress and the application of different machine learning techniques to analyze and predict stress levels. By employing methods such as KNN, Logistic Regression, Naive Bayes and Neural Network, the project demonstrated the ability to classify students into stress level categories based on their features. Additionally, Decision Trees was effective in identifying influential factors on students' stress levels.

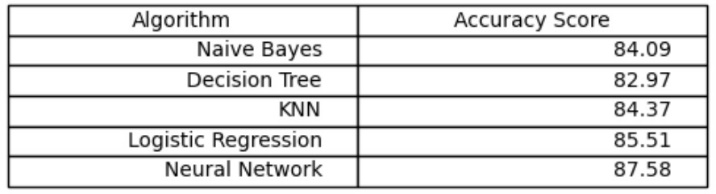


Figure 8. Algorithms and their Accuracy Score

Overall, the project provided valuable insights into student stress, its prediction, and the factors influencing it.

The outcomes of the study underscore the importance of understanding these factors for developing targeted stress management interventions. Future research could delve deeper into the causal mechanisms of these relationships and explore interventions aimed at increasing social support and academic performance to boost students' self-esteem and reduce stress levels.

Furthermore, the project emphasizes the complex interplay of various factors contributing to student stress levels, highlighting the need for an integrated approach to address student stress. This approach should encompass not only academic factors but also social support and self-esteem, reflecting the multifaceted nature of student well-being.

# Future Work

In future iterations of the project, enhancing effectiveness and improving capabilities can be achieved through various avenues. Firstly, integrating real-time data analytics can enable prompt responses to student stress by incorporating data streams from mood tracking apps or wearable devices, providing instant feedback and personalized coping strategies. This proactive approach ensures timely intervention and support, fostering a responsive and supportive environment. Additionally, personalized recommendations can tailor stress management techniques based on individual profiles, including academic performance and personal interests, acknowledging diverse experiences and coping mechanisms. Exploring sentiment analysis techniques, particularly in analyzing text data from student essays or social media, offers insights into emotional well-being, facilitating targeted interventions through advanced natural language processing algorithms. These enhancements promise to create a more adaptive and effective system for supporting student well-being.

With additional time, a more extensive exploration of hyper parameter tuning for all models could provide valuable insights into their performance. Experimenting with a broader range of parameters, particularly considering the success of Decision Trees and Random Forests opens avenues for exploring additional ensemble methods like AdaBoost or Gradient Boosting to enhance predictive accuracy. Furthermore, while the current research questions primarily focus on predicting stress levels and identifying significant contributing factors, further investigation could delve into identifying combinations of features from different domains that collectively have a greater influence on predicting stress levels. Additionally, exploring variations or ensemble approaches for the models, such as bagging or boosting techniques for Decision Trees and Random Forests, different distance metrics for KNN, and varied architectures for Multi-Layer Perceptron, presents promising directions for future research to deepen our understanding and improve the effectiveness of stress prediction models.

##### Acknowledgment

We would like to express our heartfelt gratitude to Prof., Dr. Pamela Chaudhury for her invaluable guidance and support throughout our research journey. Her expertise, encouragement, and feedback have been instrumental in shaping this work. We would also like to extend our appreciation to all the individuals who have provided assistance and support during this project. Their contributions have been essential in making this work possible. Lastly, we would like to thank our colleagues and fellow team members who have shared their insights and expertise, which have greatly enriched our understanding and contributed to the success of this project.

##### References

1. P. Manjunath , Twinkle S, P. Shreya, V. Ashok and Dr. S. Sultana. “Predictive Analysis of Student Stress Level using Machine Learning ”. 2021.
2. Qicheng Chen and Boon Giin Lee. “Deep Learning Models for Stress Analysis in University Students”. 2023.
3. Wu J, Kuan G, Lou H, Hu X, Masri MN, Sabo A, Kueh YC. “The impact of COVID-19 on students' anxiety and its clarification: a systematic review.” Front Psychol. 2023.
4. Fred Johansson, Ann Rudman, and Margreth Grotle. “Depression, anxiety and stress among Swedish university students before and during six months of the COVID-19 pandemic”. 2021.
5. https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis