Analyzing Student Stress Factors through Machine Learning Techniques

Manan Patel
Undergraduate Student
University of Tennessee
mpatel65@vols.utk.edu

Zachary Perry
Undergraduate Student
University of Tennessee
zperry4@vols.utk.edu

***Abstract*— This report aims to evaluate student stress factors and how different characteristics can contribute to these factors. Our first objective for this study is to determine if we can predict the stress level of a new student based on the similarity of their stress factors to those of known students. Secondly, we want to find which factors contribute the most significantly to predicting stress levels in students. For this report, we used the "Student Stress Factors: A Comprehensive Analysis" dataset, which contains 20 features contributing to student stress. We then leverage multiple machine learning methods, such as K-Nearest Neighbors, Multi-Layer Perceptrons, Decision Trees, and Random Forests, to determine if predicting a student's stress level is possible and which features are most important. The results of our study show that we are, in fact, able to predict a student's stress factor effectively using methods like K-Nearest Neighbors and Multi-Layer Perceptrons. Alongside this, we found that decision trees and random forests were able to find the most important, influential factors on stress, such as sleep quality, blood pressure, and safety. These findings help us better understand how to predict a student's stress level and which factors contribute the most.**

I. Introduction / Motivation

In our project, we chose a dataset that encompasses 20 features that are identified as impactful contributors to student stress. These features are categorized into Psychological, Physiological, Social, Environmental, and Academic Factors, providing a versatile analysis of student stress. We decided to focus on this dataset precisely because of its interconnection with the factors affecting student well-being.

Our project's overarching goal is to assess the feasibility of predicting new students' stress levels and identify and understand the factors that contribute most significantly to predicting stress levels in students. This involves examining the relationships between the five categories mentioned above to learn valuable insights into student well-being.

To address these goals, We utilized machine-learning approaches. To predict stress levels in new students, we employed K-Nearest Neighbors (KNN) and Multi-Layer Perceptron (MLP) models. For the identification of significant contributing factors, we employed Decision Trees and Random Forest models. For all our machine learning models, we engaged in a hyperparameter tuning process to enhance their performance.

Our success in this project will be evaluated through the models' predictive accuracy, particularly their test accuracy. Additionally, feature importance analysis will provide insights into the factors that are most impactful in predicting student stress levels. This paper proceeds into the specifics of the dataset, detailed methodologies, results, and discussions arising from our exploration.

II. Dataset

Finding an extensive and reliable dataset that fits our objectives was challenging. Many datasets we initially found often lacked features or were incredibly small in sample size. Luckily, we found the dataset "Student Stress Factors: A Comprehensive Analysis." This dataset was posted on Kaggle and included survey responses from 1,100 students aged 15 to 24 [1]. A wide age range ensured that high school and college students were accounted for and included.

The dataset provided a total of 21 features, 20 of which contributed to the overall stress factor of the individual student [1]. These factors fell into one of 5 major categories and were measured using a specific number range or a popular metric [1]. These categories include psychological, physiological, environmental, academic, and social [1].

	min	max
anxiety_level	0.0	21.0
self_esteem	0.0	30.0
mental_health_history	0.0	1.0
depression	0.0	27.0
headache	0.0	5.0
blood_pressure	1.0	3.0
sleep_quality	0.0	5.0
breathing_problem	0.0	5.0
noise_level	0.0	5.0
living_conditions	0.0	5.0
safety	0.0	5.0
basic_needs	0.0	5.0
academic_performance	0.0	5.0
study_load	0.0	5.0
teacher_student_relationship	0.0	5.0
future_career_concerns	0.0	5.0
social_support	0.0	3.0
peer_pressure	0.0	5.0
extracurricular_activities	0.0	5.0
bullying	0.0	5.0
stress_level	0.0	2.0

Fig. 1. Stress Factor Dataset Features

A. Psychological Factors

This category includes anxiety level, self-esteem, mental health history, and depression. The anxiety level of a student is measured using the GAD-7 score [1]. This scoring system

will ask students seven questions, each requiring a numerical response from 0 to 3. These responses are then added together, resulting in a final score ranging from 0 to 21 [2]. The classifications for these scores include:

1) 0-4: minimal anxiety
2) 5-9: mild anxiety
3) 10-14: moderate anxiety
4) 15-21: severe anxiety

Next, self-esteem is measured using a different scale called the Rosenberg Self-Esteem Scale [1]. This scale has a complex scoring system to determine an individual's self-esteem based on their responses to a series of questions [3]. The response to each question is a number from 4 to 1, representing strongly agree, agree, disagree, and strongly disagree, respectively. Based on the answer and the question number, a specific point value is added to the total, up to 30 [3]. Higher scores indicate higher self-esteem. Mental health history, in the case of this dataset, is measured as a boolean. So, if an individual has no history of mental health issues, they would answer with a 0. Contrary to this, a person with a mental health history would answer with a 1. Lastly, depression was measured using the Patient Health Questionnaire, or the PHQ-9. This assessment, similar to the others, asks several questions that the participants can answer with a number from 0 to 3 [4]. The final score is then calculated by adding together the numerical answers up to 27. The classifications for these scores include:

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

B. Physiological Factors

The physiological category includes the factors of headache, blood pressure, sleep quality, and breathing problems. The scaling for these features drastically differs from that of the psychological features. Here, headache, sleep quality, and breathing problems are all rated on a scale of 0 to 5 [1]. The scale is as follows:

1) 1: Low
2) 2: Medium-low
3) 3: Medium
4) 4: Medium-high
5) 5: High

As for the blood pressure feature, this was measured on a scale of 1 through 3 [1]. We found this incredibly confusing at first glace but found that the author classified the participant's actual blood pressure into one of these three categories:

1) 1: Low
2) 2: Normal
3) 3: High

C. Environmental Factors

The environmental category includes different environmental factors the students often live with. This includes noise level, living conditions, safety, and basic needs. All four of these features are measured on a scale of 0 through 5 and are classified the same as previously mentioned above [1]. Noise level is related to how loud or distracting the student's environment is. Living conditions account for and measure the quality of living of the student. Safety measures how safe they feel where they live and how safe they feel in general. Lastly, basic needs measure whether a student receives basic living needs. So, if they are receiving enough meals a day, for example.

D. Academic Factors

The academic category includes different factors within the student's academic setting that could affect their stress levels. This includes academic performance, study load, teacher-student relationship, and future career concerns. All four features are measured on the same scale, 0 through 5, as mentioned above [1]. Academic performance is a rough measure of a student's performance and school grades. Study load measures how often the students find themselves studying and their overall class load. The teacher-student relationship describes whether the student has an academic relationship with their teacher and, if so, whether it is a good or bad relationship. We initially found this to be an odd feature, but it does make sense regarding stress levels. A student with a horrible relationship with their teacher, and if their teacher is rude to them, they will probably be more stressed in that class. Lastly, future career concerns measure how worried the student is about their life after school. Again, we initially thought this was an odd feature, but it does make sense. If a student is constantly concerned with, for example, not getting a job after college, then this could definitely affect their stress levels.

E. Social Factors

The last category, social, aims to measure how social factors, both in and outside school, can affect a student's stress level. This includes social support, peer pressure, extracurricular activities, and bullying. The features are measured on the same scale, 0 through 5, except for social support [1]. Social support is measured as 0 through 3 and aims to rank the amount of support the student receives from either friends or family. This could include but is not limited to, practical help or emotional support [5]. Next, peer pressure is also measured to determine how often students are influenced by their peers to either do something or think a certain way. Extracurricular activities measure the student's level of involvement outside of school. This could include any clubs, sports, or other hobbies. Finally, bullying is the last social feature measured. It was measured to determine how often a student experiences bullying at or outside school.

All of the features in each of the five categories contribute to the student's overall stress level ranking. The scale of this ranking is from 0 to 2, classified as low, medium, and high stress levels.

We decided to keep and use all 20 features contributing to the overall stress level. In our case, all of these features are entirely possible for a student to experience and could be incredibly important in determining their stress levels. We also did not want to remove anything in case some factors ended up being more important than we initially thought. We assumed that any feature that ended up unimportant would have no effect on the stress level prediction and could just be ignored.

When first finding and analyzing the dataset, we quickly found that virtually no cleaning was needed. The 1,100 entries were all complete, with an answer for each feature included. Not having to remove any rows or substitute any NaN values was incredibly nice. There was one issue that we did initially have, and that was the scaling of each of the features. Unfortunately, when we found the dataset, there was virtually no documentation. So, the scales for the factors and what each number meant was a complete mystery to us. Luckily, after a few weeks, the dataset's author responded to a comment on Kaggle, explaining the different measurements and updating the documentation. This was incredibly lucky, and I believe that we would not have been able to use the dataset if this had not happened. While we didn't have to do much data cleaning, we did notice one thing about the data. The dataset had a slightly unbalanced number of stress level classifications. More specifically, 373, 358, and 369 students were categorized into the stress levels of 0, 1, and 2, respectively, as displayed below.

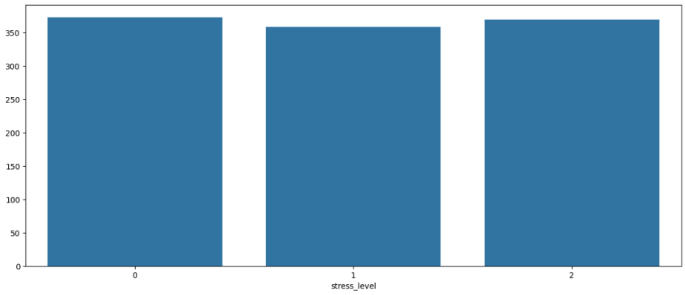

Fig. 2. Bar Chart: Number of Students in Each Stress Level Category

While we believe the balance here should not drastically affect our model's outcomes, we are interested in seeing how this could affect their performance.

With the vast number of features provided, we found this dataset intuitive and perfect for our use case. Once the documentation was updated, the feature measurement scales became much more explainable and less confusing. Overall, we believe we chose the perfect dataset for our project.

III. Machine Learning Approaches & Methodology

For our project, we decided to use four machine-learning approaches to answer both of our research questions. These approaches included using K-Nearest Neighbors, Multi-Layer Perceptrons, Decision Trees, and Random Forest. For each of them, we used a training-testing split of 80% and 20%, respectively, and always used a random state of 42 for reproducibility. Through these four methods, we would not only be able to classify new students into one of three stress level categories based on their factors but also find which features influence a student's stress level the most.

A. K-Nearest Neighbors

To begin our investigation into answering the research question," Can we predict the stress level of a new student based on the similarity of their stress factors to those of known students?" we first started with K-Nearest Neighbors, or KNN for short. This machine learning method will store the training data set, and when fed the testing data, it will find the closest training example in the dataset and classify the testing example into the same category [6]. This category will be either 0, 1, or 2, representing the different stress levels. The model was initially configured to use three neighbors. We chose to try KNN first because of its simplicity and ease of hyperparameter tuning.

The KNN model was evaluated based on its accuracy in classifying the testing data into their correct stress level categories based on the training data. The accuracy of a model is found by calculating the "accuracy_score" provided by sklearn's metric package. The results, which will be discussed more in the results section, were then plotted for future comparison.

To maximize the performance of our model, we used KFold-Cross Validation with ten splits for hyperparameter tuning and to find the most optimal number of neighbors. This involved testing all odd numbers from 1 to 30 as the number of neighbors. For each iteration, we split the training data into smaller sub-training data and small validation data sets. We would then run the KNN classifier with the current iteration as the number of neighbors, predict the validation set, and calculate the accuracy score. By doing this, we found that 27 neighbors were optimal and performed the best.

B. Multi-Layer Perceptron

Next, we used a Multi-Layer Perceptron, or MLP, classifier to again see if we could predict the stress level of a new student based on their similarity in stress factors to known students. A Multi-Layer Perceptron classifier is a type of neural network with multiple hidden layers that are all fully connected and typically perform well on classification tasks [7]. While this is a less explainable, more complex method, we wanted to try an MLP classifier to determine if its complexity, combined with its often good performance on classification tasks, would assist in outperforming KNN for the same task.

The MLP model was evaluated in the same way as the KNN model. We determine the model's accuracy by calculating the "accuracy_score" provided by sklearn's metric package. The results were then plotted similarly to the KNN results for easy comparison.

To maximize the performance of our MLP model, we decided to tune our hyperparameters. This involved scaling our data and applying K-Fold cross-validation with three

splits to find the best combination of values for the hidden layer sizes and initial learning rate parameters. We scaled our data to normalize the features that are on a larger scale. This will help improve performance and prevent the model from assigning higher importance to those features with higher scales. We first tested the hidden layer values of 50, 100, 200, and 500 in combination with the initial learning rates of 0.0001, 0.001, 0.01, 0.1, and 1. After finding the best combination, we used the initial learning rate to find the best network structure to use. This involved applying K-Fold cross-validation again but testing different network structures with our best-performing initial learning rate. Once found, we used both of these tuned hyperparameters, combined with the max number of iterations being set to 1000, to fit and calculate the accuracy of our model.

C. Decision Trees

To investigate an answer to our research question, "Which factors contribute most significantly to predicting stress levels in students?", we opted for the interpretability of Decision Trees. This machine learning approach has a hierarchical tree-like structure that systematically evaluates and selects features based on their importance in predicting the target variable – in this case, student stress levels. The Decision Tree classifier was configured with no max depth for full expansion of the tree and implemented the entropy criterion to measure information gain. The classifier was also fixated on a random state of 42 for reproducibility.

After training, the model was evaluated based on accuracy metrics to gauge its performance on unseen test data. In our study, the accuracy of a model is calculated by finding an "accuracy_score" provided by sklearn's metric package. Later, we extracted feature importance for a deeper understanding of the influential variables from the Decision Tree property, "feature_importances_"

Recognizing the importance of fine-tuning our model for optimal performance, we engaged in hyperparameter tuning, focusing on the "max_depth" parameter of the Decision Tree classifier. This iterative process involved testing several values to find out the value that maximizes the accuracy while avoiding overfitting. Subsequently, we reran the classifier with the refined parameter and documented both accuracy and feature importance.

D. Random Forest

In our exploration to uncover the key factors influencing stress levels among students, we included Random Forests. Complementing our analysis of Decision Trees, Random Forests offers a robust ensemble learning technique that leverages multiple decision trees to enhance predictive accuracy. This approach introduces an added layer of complexity by aggregating predictions from individual trees.

Much like Decision Trees, Random Forests maintain a hierarchical tree-like structure, systematically assessing and prioritizing features based on their significance in predicting our target variable–student stress levels. For our Random Forest configuration, we defaulted to 100 "n_estimators" and a random state of 42. Similar to Decision Trees, the accuracy of the model was calculated using "accuracy_score" from sklearn's metric package, and important features were noted.

We explored different values for the "n_estimators" parameter, seeking the optimal number of trees in the forest that would yield the maximum accuracy on test data. Afterward, the classifier was rerun with this refined parameter, and the resulting accuracy metrics and feature importances were documented. This iterative process enhances the model's performance, ensuring the accuracy and generalizability of the model.

IV. Results

In the Results section, we present the outcomes of our investigation into predicting and understanding stress levels among students using a diverse set of machine learning models. Addressing the research question, "Can we predict the stress level of a new student based on the similarity of their stress factors to those of known students?" we employed K-Nearest Neighbors (KNN) and Multi-Layer Perceptron (MLP) algorithms. Leveraging the inherent capabilities of KNN to identify patterns in proximity and MLP's proficiency in capturing complex relationships.

Additionally, in response to the inquiry, "Which factors contribute most significantly to predicting stress levels in students?" we turned to Decision Trees and Random Forest models. These algorithms, known for their explainibility and ensemble learning strength, provided insights into feature importance within our dataset. The subsequent analysis sheds light on the influential factors contributing to student stress and underscores the significance of hyperparameter tuning in refining the accuracy and effectiveness of our predictive models.

A. K-Nearest Neighbors

The K-Nearest Neighbors model, with its number of neighbors, tuned to 3, performed very well when predicting a student's stress level based on the similarity of their stress factors to other students. The model achieved a 92% accuracy score on the training data and an 87% accuracy score on the testing data.

While these were excellent results, we wanted to see if we could improve performance by hyperparameter tuning the number of neighbors. After applying KFold Cross-validation using all odd numbers from 1 to 30, we found that 27 neighbors performed the best on the validation set, with an accuracy of 87.6%. Using 27 neighbors, we got 87% accuracy on the training set and 89% on the testing set.

While the training accuracy did decrease when using 27 neighbors, initially, with three neighbors, the model could have been slightly overfitting, causing the training accuracy to be much higher. Also, with the 27 neighbors, the testing accuracy did increase by 2%. While not a huge difference, the model's performance did increase after this hyperparameter tuning.

One interesting thing we noticed about the model's results was which stress factor classes it kept classifying incorrectly.

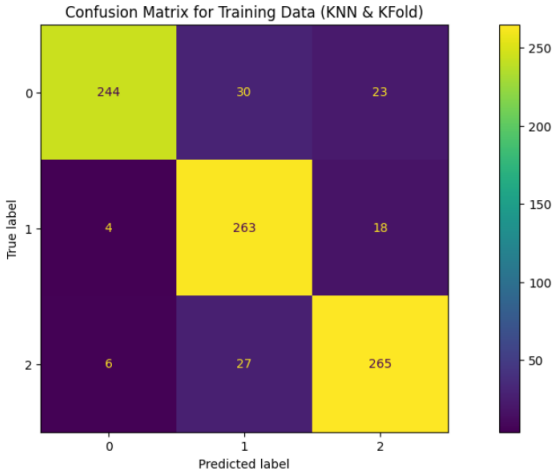

Fig. 3. Confusion Matrix: KNN Training Results

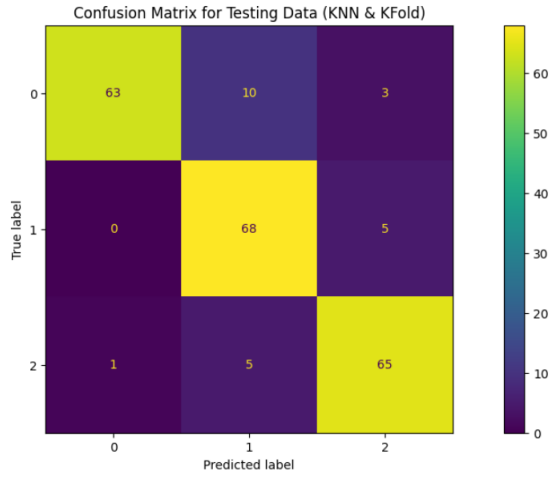

Fig. 4. Confusion Matrix: KNN Testing Results

From the figures above, it most often misclassifies the stress level of 1. During training and testing, this was the most common misclassification and was most often misclassified as a stress level 0. This could be due to the slightly imbalanced dataset and the fact that there were more students within the stress level zero category than in the stress level 1 category. To find out, we undersampled the number of students in both the stress level 0 and 2 categories to see if this would improve performance.

After undersampling stress levels 0 and 2 and applying KFold cross-validation, we found that one singular neighbor was the best-performing number of neighbors on the validation set. Then, when applying this to the undersampled training and testing data, we got 100% training accuracy and almost 87% testing accuracy. While the training accuracy improved, this could be due to overfitting since only one neighbor was used for the KNN model. Alongside this, the testing accuracy was slightly lower than before, proving that undersampling had little effect on the outcome.

B. Multi-Layer Perceptron

We started with an initial learning rate of 0.0001 and 50 hidden layers as the parameters for the multi-layer perceptron model. This would give us a good baseline starting point for performance. The model produced the same accuracy for both the testing and training sets with these two parameters. This accuracy was 88%.

After this initial baseline test, we deployed KFold cross-validation to start tuning the hyperparameters. After testing numerous combinations of hidden layer values and initial learning rate values, we found the best-performing pair with an initial learning rate of 0.01 with 500 hidden layers as seen in Figure 5.

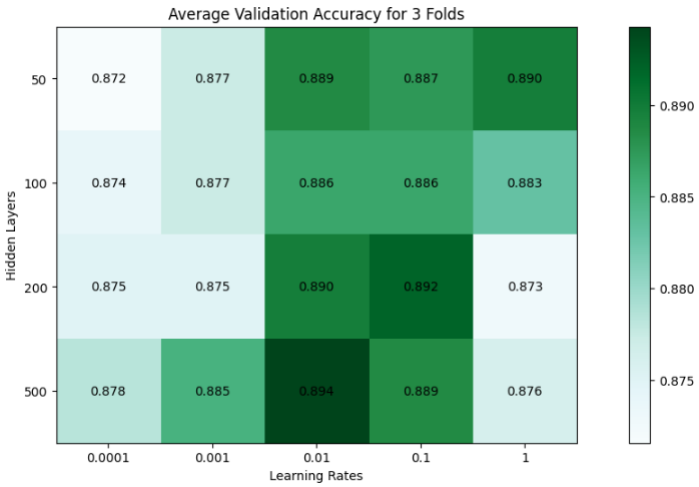

Fig. 5. Heatmap: MLP Hyperparameters

We tested many different network structures to confirm the number of hidden layers further. For example, we tested [100], [100, 100], [500], [500, 200], and so on. We found that simply using 500 hidden layers performed best on the validation set. With both of these parameters, we found the training and testing accuracies for the MLP classifier. For training, it earned a perfect score of 100%. For testing, it earned an accuracy score of 88%.

While the MLP classifier did perform well, it still did not perform as well as KNN did on the testing set. We also noticed that the model was doing the same thing as KNN, often misclassifying stress levels of 1 as 0.

With a similar suspicion, we decided to try undersampling to better balance the number of students classified into each stress level category. Applying the same hyperparameter tuning, we found that the best-performing initial learning rate was 0.1, and the best network structure was (100,100). The training accuracy was, again, 100%, but the testing accuracy was 87.9%, just under 88%.

Undersampling did not help improve the performance of the MLP model, and, ultimately, it performed worse than KNN on the dataset. We believe that if we had potentially more data, the model would perform better. Alongside this, the model could have been looking for more complex relationships between features when there weren't any.

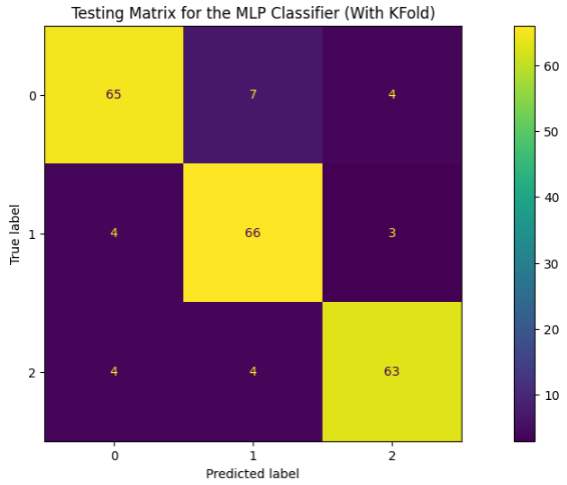

Fig. 6. Confusion Matrix: MLP Testing Results

C. Decision Trees

The predictive model, using a Decision Tree with maximum depth, demonstrated an outstanding performance in predicting stress levels among students. The model achieved a training accuracy of 100%, indicating its ability to capture the underlying patterns within the training data. Upon evaluation of the model on the test data, it got an accuracy of 88%, highlighting the generalization capabilities.

In terms of feature importance, the findings from Figure 7 revealed that physiological and social factors play crucial roles in predicting stress levels. Blood pressure turned out to be the most influential feature, contributing significantly (40%) to the predictive accuracy of the model. Subsequently, social support demonstrated 29% importance, emphasizing the impact of interpersonal relationships on student stress.

Other noticeable contributors include sleep quality (9.81%), bullying (4.37%), and breathing problems (4.03%). These findings align with existing research on stress, highlighting the relationship between physical health, social dynamics, and emotional well-being in the student population.

On the contrary, certain features such as mental health history, headache, and basic needs showed minimal importance (0%). While these factors may not be insignificant in the broader context of student well-being, their limited impact in predicting stress levels suggests that the model prioritizes other features in its decision-making process.

The Decision Tree model, after thorough hyperparameter tuning, demonstrated robust performance with a training accuracy of 89.0% and a test accuracy of 90.0%. The hyperparameter tuning involved testing various values for the "max_depth" parameter, and the optimal value was determined to be 3, as shown in Figure 8. This choice reflects a careful balance, avoiding overfitting (as evidenced by the perfect train accuracy) while maintaining a high level of generalization, as indicated by the 90.0% test accuracy.

As demonstrated in Figure 9, the most critical features remain as mentioned above, blood pressure, social support, and sleep quality.

Rank	Feature	Importance
1	blood_pressure	40.21%
2	social_support	29.07%
3	sleep_quality	9.81%
4	bullying	4.37%
5	breathing_problem	4.03%
6	academic_performance	2.12%
7	depression	2.09%
8	extracurricular_activities	1.97%
9	teacher_student_relationship	1.51%
10	peer_pressure	0.88%
11	self_esteem	0.84%
12	noise_level	0.81%
13	study_load	0.70%
14	anxiety_level	0.70%
15	safety	0.34%
16	future_career_concerns	0.23%
17	living_conditions	0.20%
18	basic_needs	0.14%
19	headache	0.00%
20	mental_health_history	0.00%

Fig. 7. Max depth Decision Tree feature importances

D. Random Forest

Random Forest classifier demonstrated a training accuracy of 100% and an admirable test accuracy of 89%. This indicates that the model is capable of understanding the data very well and its ability to generalize on unseen data judging by its 89% test accuracy.

To optimize the Random Forest model, we performed a hyperparameter tuning process (Figure 10). Specifically, we tested various values of the "n_estimators" parameter of the classifier, discovering that the optimal number of trees in the forest was 20. This tuning process enhances the model's performance by preventing overfitting and promoting better generalization to unseen data.

The Random Forest model's feature importance analysis (Figure 11) unveiled the key contributors to predicting student stress levels. Some notable features include sleep quality (14.46%), blood pressure (12.17%), and safety (8.62%) in the order they are mentioned. These features were identified to be the most impactful predictors of stress levels. The

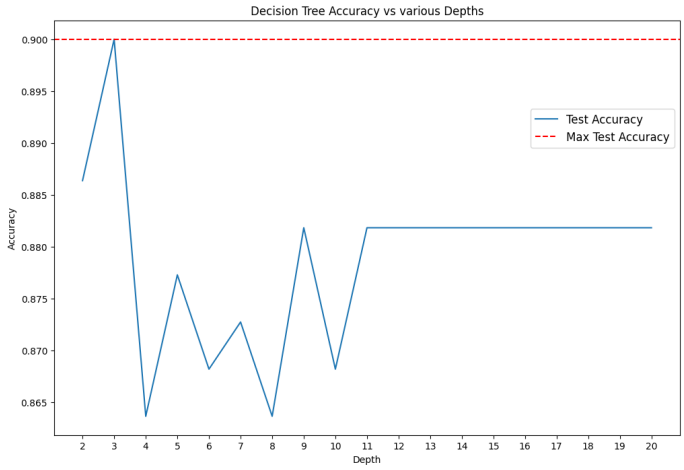

Fig. 8. Compare several depths of Decision Tree

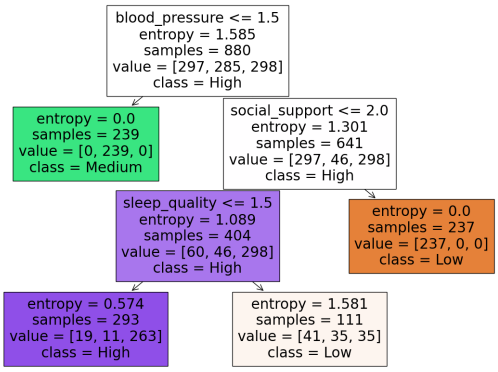

Fig. 9. Depth-3 Decision Tree Feature Importances

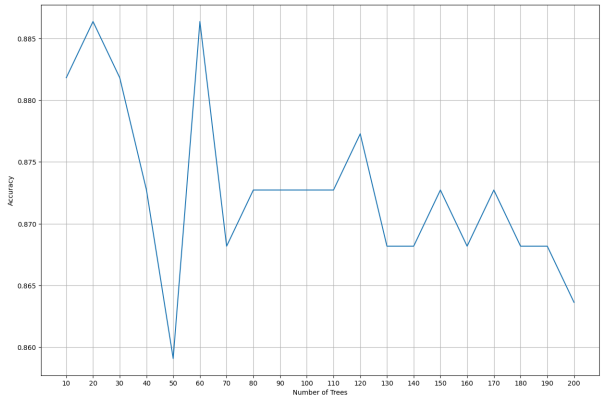

Fig. 10. Testing n_estimators for Random Forest classifier

Rank	Feature	Importance
1	sleep_quality	14.46%
2	blood_pressure	12.17%
3	safety	8.62%
4	basic_needs	8.11%
5	self_esteem	6.96%
6	anxiety_level	6.51%
7	teacher_student_relationship	6.40%
8	extracurricular_activities	6.17%
9	bullying	5.11%
10	headache	4.68%
11	noise_level	3.71%
12	social_support	3.59%
13	depression	3.53%
14	academic_performance	3.27%
15	study_load	1.43%
16	peer_pressure	1.40%
17	future_career_concerns	1.32%
18	living_conditions	1.15%
19	breathing_problem	0.79%
20	mental_health_history	0.61%

Fig. 11. Hyperparameter tuned Random Forest classifier feature importances

remaining features, such as teacher-student relationship, self-esteem, anxiety level, and extracurricular activities, also contribute to the model's predictive power, each capturing different aspects of the student experience.

V. Discussion

After applying K-Nearest Neighbors and Multi-Layer Perceptrons, we achieved high accuracies when predicting the stress level of a new student based on the similarity of their stress factors to those of known students. We found that hyperparameter tuning did, indeed, improve model performance overall. We found that K-Nearest Neighbors performed best, with an overall 89% accuracy rating for the testing set. Although K-Nearest Neighbors performed better, it was not by a considerable margin, as the Multi-Layer Perceptron model still performed with an 88% accuracy rating on the testing set.

Overall, we can confidently say that K-Nearest Neighbors is the better model for classifying a student into a specific stress level based on their stress factors and the stress factors of others. Not only does it have a higher testing accuracy, but it also makes more sense to use in our case. With a classification problem such as ours, we are trying to classify students into one of three stress levels. Applying KNN to this task allows for a more explainable, more straightforward approach than using a more complex model such as MLP.

The feature importance derived from the Decision Tree and Random Forest models provides hierarchical significance of various factors in predicting student stress levels. In both models, physiological factors, such as blood pressure and sleep quality, emerged as the most impactful variables. Blood pressure was found to be the most crucial feature, indicating the link between physical health and stress. Similarly, sleep quality, played a significant role in both models, confirming the well-documented relationship between inadequate sleep patterns and increased stress among students.

Additionally, the importance of social factors, such as social support, also ranked high in terms of impactful features in both models. This supports the importance of interpersonal relationships in mitigating or intensifying stress among students. The emphasis given to social support is backed by existing studies, which indicate the role of strong social connections in reducing stress and cherishing emotional well-being.

The comparable performance of the Decision Tree and Random Forest models in predicting stress levels showcases

the robustness of the Decision Tree model despite its simplicity compared to the ensemble approach of Random Forest.

VI. Conclusion

This report comprehensively evaluated key categories of factors contributing the most towards student stress. By applying different machine learning techniques, we were able to analyze student stress factors further and how they contribute to an overall stress level. We found that applying methods such as K-Nearest Neighbors and Multi-Layer Perceptrons would allow for the functionality of classifying a new student into one of three stress level categories based on their other features. We also found that Decision Trees and Random Forest approaches were able to find the most influential factors on a student's stress level. Overall, applying these different machine-learning methods provided valuable insights into student stress, how we can predict it, and what influences it the most.

VII. Future Work

With additional time, we might consider conducting a more extensive exploration of hyperparameter tuning for all our models. Experimenting with a broader range of parameters could provide additional insights into the model's performance. Considering the success of Decision Trees and Random Forests, exploring additional ensemble methods like AdaBoost or Gradient Boosting could provide insights into whether combining multiple weak learners could enhance predictive accuracy.

While the current research questions focus on predicting stress levels and identifying significant contributing factors, the exploration of additional questions could arise based on the findings. For example, are there certain combinations of features from different domains (Psychological, Physiological, Social, Environmental, Academic) that collectively have more influence on predicting stress levels?

Some aspects weren't fully explored due to time constraints. For instance, We could explore the variations or ensemble approaches for the models explored in this study. For example, implementing bagging or boosting techniques for Decision Trees and Random Forests, different distance metrics for KNN and varied architectures for Multi-Layer Perceptrons might yield diverse perspectives on stress prediction.

VIII. Contribution of Team Members

A. Zachary Perry

- Focused on predicting stress levels of new students based on similarity to known students.
- Conducted in-depth research and testing of KNN and MLP models and hyperparameter tuning of each model for optimal performance
- Created appropriate visualizations

B. Manan Patel

- Worked on answering the factors that contributed significantly to predicting stress levels in students
- Explored Decision Trees and Random Forest models, conducting thorough testing and hyperparameter tuning for each
- Created appropriate visualizations

C. Both

- Collaboratively discovered and selected the dataset for the project
- Formulated the research questions jointly
- Shared responsibilities in creating a presentation and co-authored the paper

References

[1] "Student Stress Factors: A Comprehensive Analysis," www.kaggle.com. https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis

[2] R. Spitzer, "GAD-7 Anxiety," 1999. Available: https://adaa.org/sites/default/files/GAD-7_Anxiety-updated_0.pdf

[3] M. Rosenberg, "ROSENBERG SELF-ESTEEM SCALE," fetzer.org, 1965. Available: https://fetzer.org/sites/default/files/images/stories/pdf/selfmeasures\/Self_Measures_for_Self-Esteem_ROSENBERG_SELF-ESTEEM.pdf

[4] "PATIENT HEALTH QUESTIONNAIRE (PHQ-9)," Oct. 2005. Available: https://med.stanford.edu/fastlab/research/imapp/msrs/_jcr_content/main/accordion/accordion_content3/download_256324296/file.res/PHQ9%20id%20date%2008.03.pdf

[5] American Psychological Association, "APA Dictionary of Psychology," dictionary.apa.org, 2020. https://dictionary.apa.org/social-support

[6] C. Schuman. NearestNeighbors [PowerPoint slides]. University of Tennessee, Knoxville, 2023. Available: https://utk.instructure.com/courses/179872/files/17767847?module_item_id=3771387

[7] C. Schuman. TeamProject_NeuralNetworks_Part2 [PowerPoint slides]. University of Tennessee, Knoxville, 2023. Available: https://utk.instructure.com/courses/179872/files/18465674?module_item_id=3820936