*Course Number: CMPS320 Instructor: Prof. Tweneboah*

**Analysis and Prediction of Stress Levels Based on Students’ Stress Factors**

**Tiffany LaTorre 1, Tina Nosrati 2, Kamila Palys 3, and Jenna Vesey 4**

1 Ramapo College of New Jersey 1; [tlatorre@ramapo.edu](mailto:tlatorre@ramapo.edu)

2 Ramapo College of New Jersey 2; [tnosrati@ramapo.edu](mailto:tnosrati@ramapo.edu)

3 Ramapo College of New Jersey 3; [kpalys@ramapo.edu](mailto:kpalys@ramapo.edu)

4 Ramapo College of New Jersey 4;[jvesey@ramapo.edu](mailto:jvesey@ramapo.edu)

**Abstract:** With stress being one of the main concerns regarding students' academic performance, researchers have been focused on gathering and analyzing data regarding the roots of the problem and finding the best approach to address it. In this research project, we focused on data collected from the students' stress factors in Nepal to investigate the most indicative of students' stress levels. Using classification techniques, we also developed a predictive model for students' stress levels based on these factors. Our findings demonstrate that although no specific group of variables, such as physical, psychological, or social factors, contribute to students' stress, a combination of factors from each group plays a crucial role in predicting stress. In conclusion, we recommend that institutions focus their policies on improving a group of recommended factors such as students' sleep quality, controlling bullying, and student-teacher relations.

**Keywords:** Stress Factors; Students Performance; Stress Prediction; Stress Level Classification; Cluster Analysis;

**1. Introduction**

Stress is how the body deals with environmental changes [1]. This stress manifests itself as emotional, physical, or behavioral change. It affects everyone, including students. This can be a short-term reaction due to upcoming deadlines, but sometimes, stress can be due to traumatic or ongoing events. Stress can motivate students to do their tasks and perform better, but too much stress can harm the student. If these problems are not addressed, it leads to other issues for the student [1]. Students in high schools may end up isolating themselves due to stress. They could have outbursts when frustrated or develop harmful behaviors to deal with the stress. This chronic stress may lead to a variety of mental health conditions [2]. Stress symptoms can manifest in behavioral changes and mood changes and even manifest into symptoms. These changes can include fatigue, headache, sleep problems, memory problems, lack of motivation, and other symptoms. The human body is not meant to undergo constant stress.

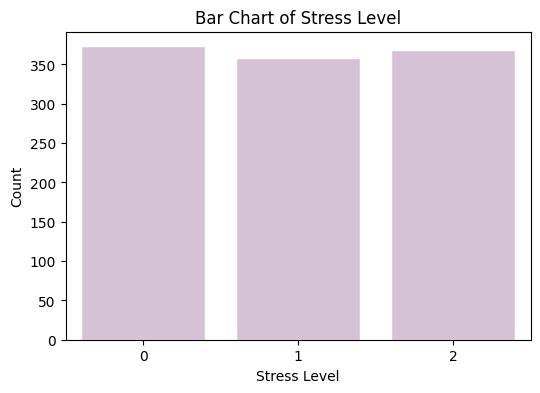
Recently, there has been research into students and how much stress they have been undergoing. Two studies have found that stress and anxiety have increased substantially since the 1950s [3]. This increase was so much more that school children in the 1980s experienced as much stress as psychiatric patients did in the 1950s. These studies have demonstrated that the number of students being diagnosed with cases of depression is going to increase in the coming decades. Because of these mental health issues arising from stress, other issues are on the rise as well [3]. Studies have suggested that alcohol and drug abuse will also continue to be an increasing problem. This increase in stress demonstrates sufficient evidence that student mental health is worsening. A study from Healthy Minds found that during the 2020-2021 school year, more than 60% of college students met the criteria to receive a diagnosis for at least one mental health issue [4]. Another American College Health Association survey indicated that almost 75% of students experience moderate to severe stress levels. Studies have shown that more students are seeking mental health resources on campus. Between 2009 and 2015, 40% more students needed mental health counseling. This increase in caseload is not proportional to the amount of funding these departments receive. This lack of funding has called for increased creative resources and strategies to limit students' stress. This need to understand the basis of stress could lend itself to creating these resources and eventually aim to lower student stress levels.

**2. Materials and Methods**

*2.1. Data Exploration*

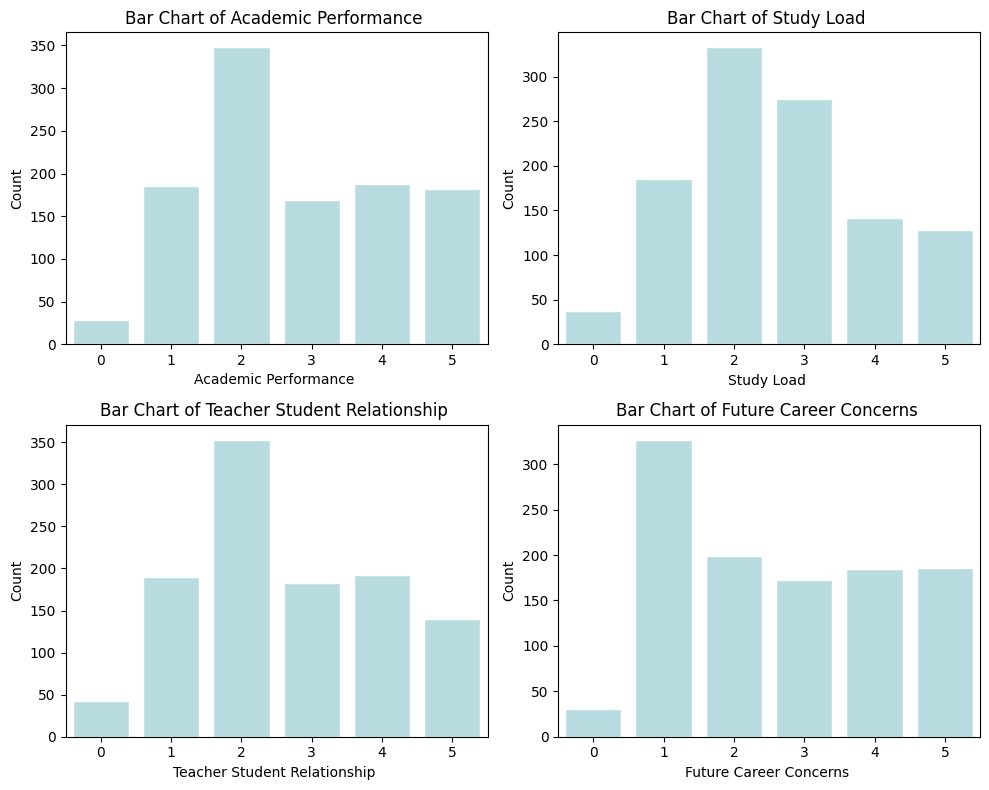
The dataset used for this project can be found on Kaggle [5]. It contains 1100 observations representing students and 21 variables representing various personal ratings the students each specified for themselves. The target variable in the dataset is stress level. Each predictor variable falls into a specific category of stress factors, with the categories being social, academic, psychological, physiological, or environmental. Most of these variables are on a scale of 0-5, with a few having a smaller or larger scale, making all of them ordinal variables. Each of these factors indicates specific levels for the participant. 0-1 indicated a low level of that factor, 2-3 indicated a moderate level, and 4-5 indicated a high level of that factor. Take safety as an example. Someone who indicated a one on safety would feel as though they had little safety, while someone with a four would feel relatively safe. These five variables are split up into a list for each. The complete list of predictor variables includes anxiety level, self-esteem, mental health history, depression, headache, blood pressure, sleep quality, breathing problem, noise level, living conditions, safety, basic needs, academic performance, study load, teacher-student relationship, future career concerns, social support, peer pressure, extracurricular activities, and bullying.

Before starting the modeling process, each variable was explored with a bar chart displaying the counts of each variable’s values. As seen in Figure 1, the target variable, stress level, has evenly distributed classes, with 373 students claiming a stress level of 0, 353 having a stress level of 1, and 369 having a stress level of 2.



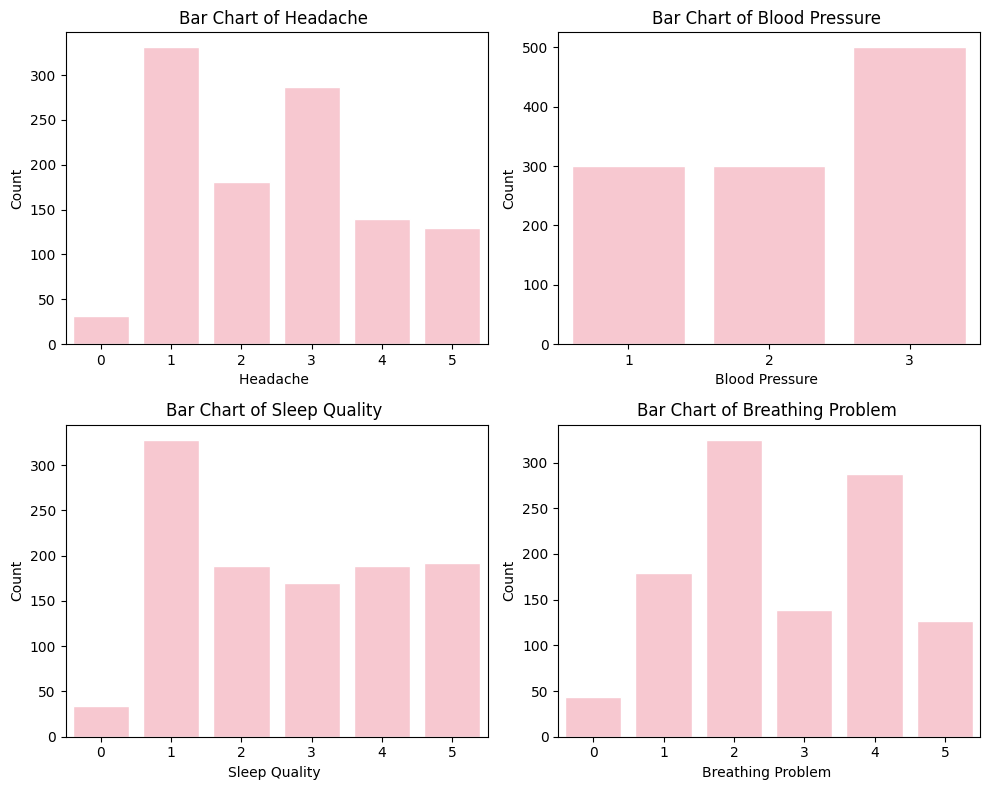
**Figure 1**

**Figure 1.** Bar chart of Stress Levels.



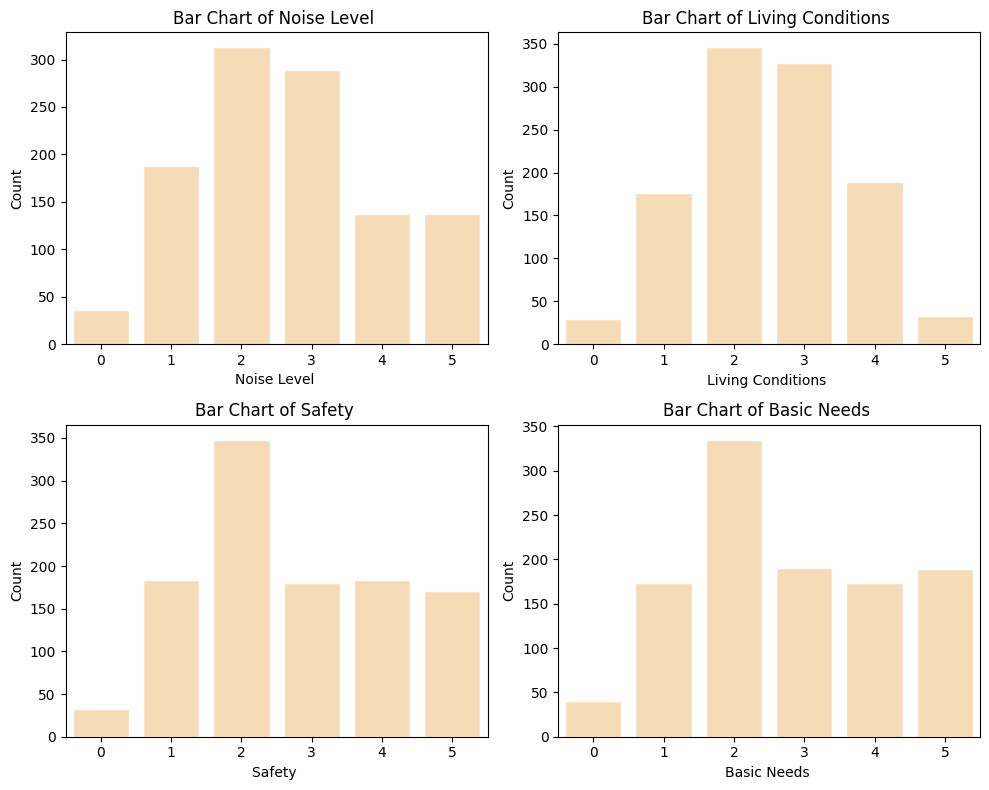
**Figure 2**

**Figure 2.** Bar charts of Academic Factors and the distribution of responses.



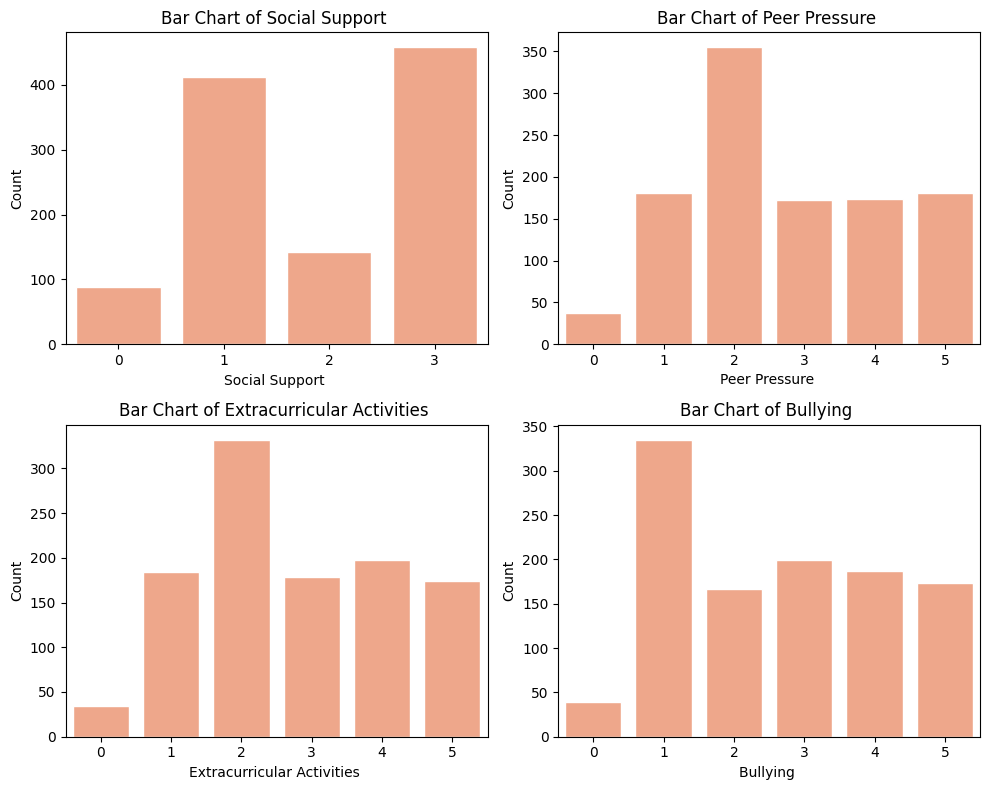
**Figure 3**

**Figure 3.** Bar charts of Physiological Factors and the distribution of responses.



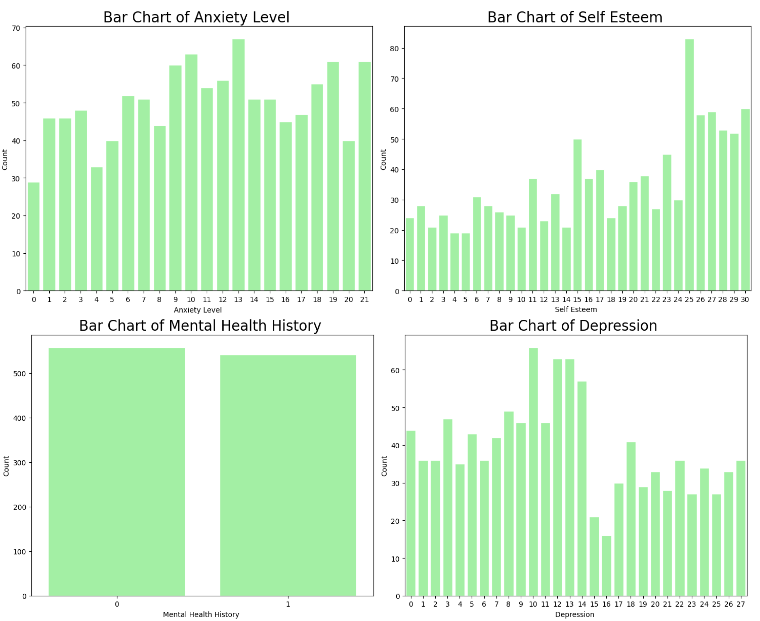
**Figure 4**

**Figure 4.** Bar charts of Environmental Factors and the distribution of responses.



**Figure 5**

**Figure 5.** Bar charts of Social Factors and the distribution of responses.



**Figure 6**

**Figure 6.** Bar charts of Psychological Factors and the distribution of responses.

Data exploration revealed deep patterns in various features in the data set. For academic factors in Figure 2, most respondents assessed their academic performance, study load, and teacher-student relations at a moderate level. Meanwhile, future career concerns were much lower, indicating little worry among the individuals surveyed about their future professional paths. As for physiological factors in Figure 3, blood pressure measurements are categorized on a scale of 1 to 3, where 1 represents low blood pressure, 2 is normal, and 3 indicates high blood pressure. Participants reported low headaches and low-rated sleep quality but higher blood pressure levels and moderate breathing problems.

Regarding environmental factors, participants reported moderate levels across noise levels, living conditions, safety, and meeting basic needs, as shown in Figure 4. Socially, the dataset highlighted a robust support system, with social support ratings notably high in Figure 5. At the same time, peer pressure and engagement in extracurricular activities were reported moderately, and bullying was relatively low. In terms of psychological factors, shown in Figure 6, the majority reported a moderate level of anxiety based on the GAD-7 scale [6]. Meanwhile, self-esteem, measured by the Rosenberg Self-Esteem Scale [7], was high.

Interestingly, most respondents had no mental health history, and there was no significant difference between those with or without such a history. Depression was surveyed, and most students scored to have moderate depression based on the PHQ-9 scale [8]. This thorough exploration provides insight into surveyed students, facilitating a comprehensive understanding of their academic, physiological, environmental, social, and psychological experiences and how stress may or may not influence them.

*2.2. Clustering*

In order to understand the hidden pattern in our dataset and investigate the relation between our variables, we decided to put our target variable aside and approach the problem as an unsupervised problem by only looking at the predictors and developing a K-Mode clustering model with 3 clusters to divide our data. Our observations showed that these clusters significantly differ in terms of stress factor levels. Comparing clusters with the actual label of the observations indicates that 90% of the observations in cluster 0 were actually labeled as zero. This number is 80% for Level 1 and 95% for Level 2.

*2.3. Classification and Predictive Model*

All classifiers were created using the sci-kit-learn library. Data was randomly split into training and testing sets using an 80/20 split. The decision tree classifier was the first model developed. We tuned the model's hyperparameters through an exhaustive grid search, covering criteria such as Gini, entropy, and splitter types (best, random), maximum depth, minimum samples for splitting, minimum samples per leaf, and maximum features. Employing the GridSearchCV function with 5-fold cross-validation and accuracy scoring facilitated the identification of optimal hyperparameters. The refined decision tree classifier used the best hyperparameters on the designated training set. The model's performance was assessed on an independent test set, and the final accuracy scores and the best hyperparameters were outputted.

Next, a random forest classifier is created using all predictor variables. Again, cross-validation is performed with this classifier to determine the best set of hyperparameters for the model. Then, each feature is output along with its importance score in this model, sorted in descending order. This allows us to answer which factors and categories of factors are typically most critical in contributing to stress levels in students. However, each time the classifier is trained, the set of most important features slightly changes due to the random component of the classifier. In addition, the output is used to perform feature selection. The feature importances are computed 20 times, each with a new random forest classifier, and the average feature importances are computed for each feature. Since there were initially 20 predictor variables, any variable with an average importance score above 0.05 is kept in the model. Only these predictor variables are used in the next random forest classifier. The accuracy score is computed for the best version of this classifier. This performance metric is chosen since stress level classes are evenly distributed, and it can also tell how well we are able to predict student stress levels.

Subsequently, the Support Vector Machine (SVM) model was developed using the SVC class. The model underwent hyperparameter tuning through a grid search, exploring C, gamma, and kernel types (linear, rbf, poly). The best hyperparameters were identified using GridSearchCV with 5-fold cross-validation and accuracy scoring. The refined SVM model, incorporating the optimal hyperparameters, was then evaluated on an independent test set, and the accuracy of the test set was calculated.

The next model to be developed was the Adaboost. This model was developed using the Adaboost algorithm. In the AdaBoost algorithm, a boosted classifier is trained. The model underwent hyperparameter tuning through a grid search, exploring learning rates and estimators. The best learning rate was determined to be 1.0, and the best number of estimators was 50. The best hyperparameters were identified using GridSearchCV with 5-fold cross-validation and accuracy scoring. The refined AdaBoost model, incorporating the optimal hyperparameters, was then evaluated on an independent test set, and the accuracy of the test set was calculated. The second to last model designed was a bagging model. In this model, our AdaBoost model was applied to our best decision tree model. Since the two models that were used within our bagging model had the best hyperparameters set, this model will also be optimized. This model had its accuracy calculated as well.

The final model designed for this analysis was a voting classifier model. In this classification model, we implemented hard voting as the way to determine the classifier. The hard voting classifier aggregated the predictions of each of our best models determined in the previous step and combined them into one. This model was then checked for accuracy.

**3. Results**

*3.1 Model Performance Metrics*

3.1.1 Decision Tree Classifier

* The decision tree classifier was trained using specific hyperparameters found using a 5-fold cross-validation model.
* The best hyperparameters are as listed: 'criterion': 'entropy,' 'max\_depth':None, 'max\_features': 'log2', 'min\_samples\_leaf': 2, 'min\_samples\_split': 2, 'splitter': 'random'
* This model achieved an accuracy of 85% with multiple iterations performed. This indicates a relatively high accuracy on our model.

3.1.2 Random Forest Classifier

* The Random Forest Classifier was trained using hyperparameters obtained using a 5-fold cross-validation model.
* The best hyperparameters are as listed: 'criterion': 'entropy,' 'max\_depth':None, 'max\_features': 'log2', 'min\_samples\_leaf': 2, 'min\_samples\_split': 2, 'splitter': 'random.'
* The Random Forest classifier achieved an accuracy of approximately 86% on the test set. This indicates a relatively high accuracy on our model.

3.1.3 Support Vector Machine (SVM) Model

* The Support Vector Machine (SVM) Classifier was trained using hyperparameters obtained using a 5-fold cross-validation model.
* The best hyperparameters used for the experiment include: 'C': 100, 'gamma': 1, 'kernel': 'linear.'
* The Support Vector Machine (SVM) Model achieved an accuracy of approximately 87%. This indicates a relatively high accuracy of our model.

3.1.4 AdaBoost Model

* The AdaBoost Classifier was tested using various n estimators and learning rates. These testing included 'n\_estimators': [50, 100, 200] and 'learning\_rate': [0.01, 0.1, 1.0].
* The best parameters found for this model and were used to determine the accuracy score are as listed: 'learning\_rate': 1.0, 'n\_estimators': 50.
* The AdaBoost Model achieved an accuracy of approximately 88%. This indicates a relatively high accuracy of our model.

3.1.5 Bagging Model

* The Bagging Classifier was created using our best Random Forest Classifier model with the AdaBoost Model applied to it.
* The Bagging Model achieved an accuracy of approximately 87%. This indicates a relatively high accuracy of our model.

3.1.6 Voting Classifier Model

* The Voting classifier model was created by combining the best Random Forest model, the best AdaBoost Model, the best Decision Tree, the best Support Vector Machine model, and the best Bagging Classifier model. This was combined using a hard voting method.
* The hard voting method allows the classifier with the most votes to aggregate into the classifier.
* This Voting Classifier Model achieved an accuracy of approximately 88%.

3.1.7 Identifying the Best Classifier Model

* The best classifier model would be the one with the greatest accuracy score.
* As seen in Figure 7, the AdaBoost Model is the most accurate model, with an accuracy score of 87%, and is also considered the best model.

*3.2 Feature Importance Analysis*

3.2.1 Determining Feature Importance

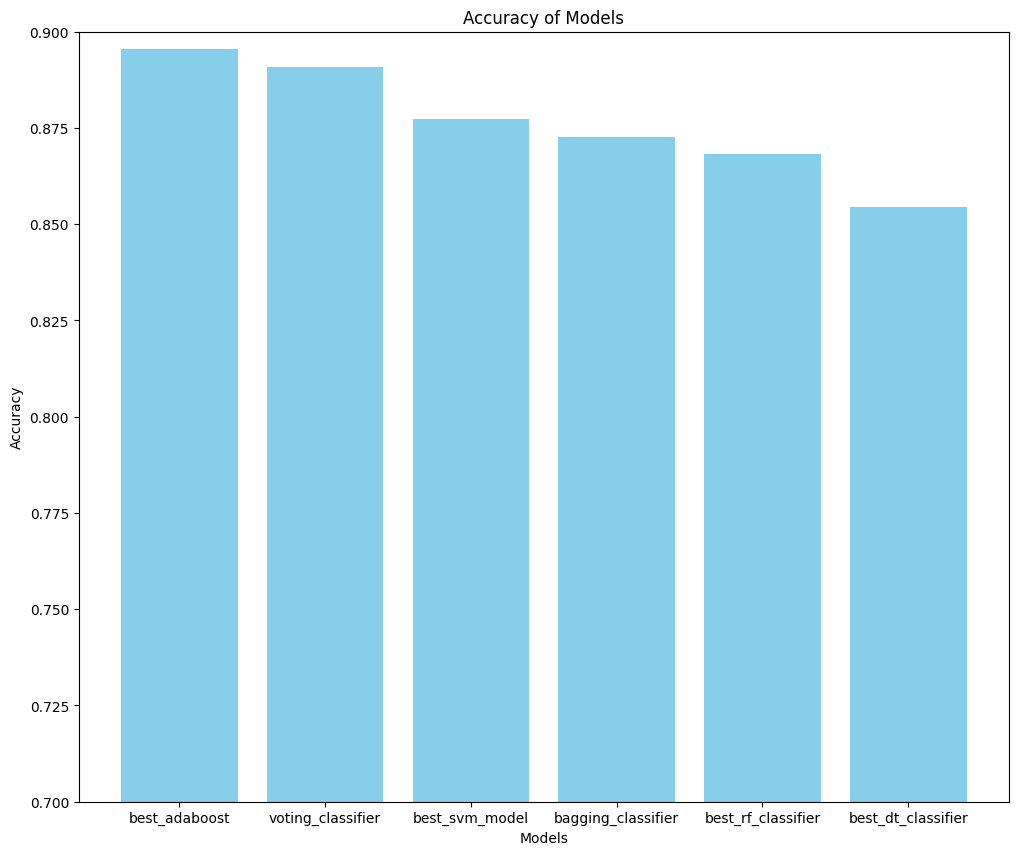
* After averaging each feature’s importance score as a result of 20 iterations of the best Random Forest classifier, we identified eleven variables with a mean importance score of greater than 0.05 (not listed in order):

1. Blood Pressure (Physiological)
2. Sleep Quality (Physiological)
3. Extracurricular activities (Social)
4. Safety (Environmental)
5. Bullying (Social)
6. Teacher Student Relationship (Academic)
7. Blood Pressure (Physiological)
8. Extracurricular activities (Social)
9. Safety (Environmental)
10. Bullying (Social)
11. Teacher Student Relationship (Academic)

* These variables are subject to change due to the random nature of the Random Forest Classification model.

3.2.2 Model Refinement and Accuracy

* A new random forest model was optimized using only these eleven variables. This random forest model was optimized using the hyperparameters found in the previous random forest.
* The best hyperparameters are as listed: 'criterion': 'entropy,' 'max\_depth':None, 'max\_features': 'log2', 'min\_samples\_leaf': 2, 'min\_samples\_split': 2, 'splitter': 'random.'
* This Random Forest model achieved an accuracy of approximately 86%. This accuracy score indicates it is relatively accurate in determining stress levels.
* Notably, this is not a better model than the Random Forest model with all variables versus the top eleven variables.



**Figure 7**

**Figure 7.** Bar chart displays the accuracy of each model.

**4. Discussion**

With the reduced set of features, the Decision Tree classifier had a high classification accuracy of approximately 86%. Removing nine out of twenty predictor variables was likely a good choice because it significantly reduced the complexity of the model while still resulting in high model performance. However, the decision tree model does not have the highest accuracy score. Although useful, it may not be the best model to obtain the desired results. This model is one of the few applicable to ordinal variables, making it worth exploring. The Random Forest model had a high classification accuracy of approximately 86%. Working on the Random Forest model using the top eleven variables produced almost the same accuracy score while simplifying the model. This indicates that not all variables are required to make an accurate model. Some variables are weighed more than others, indicating that some contribute more to stress. The Support Vector Machine (SVM) model had a high classification accuracy of approximately 86%. This model does not have the highest accuracy score either. The Bagging model had a high classification accuracy of approximately 87%. This indicates that the model could be useful for determining stress factors. This model applies to ordinal variables, such as the data in our dataset. The voting classifier also has a high classification accuracy of approximately 87%. This model is similarly useful for determining stress variables. It has the second-highest accuracy of all models, demonstrating it as a viable source of information.

The AdaBoost model is the best-performing model, with an accuracy of 89%. The AdaBoost model's high accuracy indicates that it effectively predicts students' stress levels. The performance of the AdaBoost model potentially lends itself to practical applications. A function created takes a user's input of the values of their stress factors and returns the prediction of their stress level using the AdaBoost model. This model could potentially facilitate early intervention on students by rating their stress scores and providing a way to identify targets for early interventions.

These eleven variables offer important insight into targets for understanding students' stress levels. However, there is acknowledgment that some of these variables could be a symptom of increased stress, not necessarily the cause. Blood pressure, for instance, is a symptom of individuals who have undergone high levels of stress for a substantial period [9]. Therefore, the importance of this feature should not be blindly interpreted as a predictor or cause of stress. Some variables can be considered targets for schools to implement methods to lower stress levels. Bullying and Teacher-student relationships could be areas to focus on in schools in order to combat the rise in stress. However, it is important to note that the feature importances that result from the random forest model, as the name suggests, are based on some randomness. This leads to inconsistencies in which of our variables will be over the 0.05 threshold. There may be variables over the iterations that sometimes end up on the list and are not on the list other times. Still, these are the eleven variables we see most often when the variables' importance is averaged. It has also been seen that no specific category of variables is more significant in predicting stress levels. No category was significantly more represented than another for the eleven variables identified. The highest was social factors, contributing to 3/11 of the top feature importances, while the rest tied for 2/11. This demonstrates that all the categories most likely play an equally important role in the contribution of stress and that not one factor alone leads to more stress in students. This demonstrates how multifaceted stress is based on many factors within a student's life. Stress is a diverse circumstance and needs to be approached as such within schools.

**5. Conclusions**

The concrete conclusions we can make from this study are limited because each time a classifier is initialized and trained, there will be slightly different results because of the random component in our models. However, our analysis still allows us to derive some inferences. All of the classifiers produced high accuracy scores of above 80%, which answers our research question of how well we are able to predict student stress levels based on their stress factors. Most of the time, our model correctly predicts a student’s stress. In addition, while the most important features also vary with each iteration of the random forest classifier, the eleven features we used in the final models consistently made the list of most important features. This information could be used by schools to help appropriately allocate funds to create resources for students to help manage their stress levels. Additionally, we found that no particular category of stress factors was most responsible for contributing to student stress. The categories that the eleven most important stress factors belong to are evenly distributed, with each category being represented in the feature list two to three times.

Limitations accompanied our findings as well. One of them was the fact that all the features were ordinal, meaning that the methods we could use in our data exploration and creating models were limited. In addition, the inherent randomness of the random forest classifier made it challenging to designate a set of variables as being most important. Because of this, the best we could do was to run the classifier multiple times and see the most common results. This renders our conclusions susceptible to change and not definitive. Finally, the dataset concerns students specifically in Nepal, so our findings may not be extended to other populations. Each country has a distinct culture and values with different stress factors that can contribute to a student’s stress level. Data must be collected from other countries in order to be able to make conclusions pertaining to the students there.

**Author Contributions:** Conceptualization, Tiffany LaTorre, Tina Nosrati, Kamila Palys, and Jenna Vesey.; Methodology, Tiffany LaTorre, and Tina Nosrati.; Software, Tina Nosrati, Tiffany LaTorre, Kamila Palys, and Jenna Vesey.; validation Tina Nosrati, Kamila Palys., Jenna Vesey, and Tiffany LaTorre.; formal analysis, Tina Nosrati.; data curation, Tina Nosrati, Kamila Palys, Jenna Vesey, and Tiffany LaTorre.; writing—original draft preparation, Jenna Vesey, Tina Nosrati, Tiffany LaTorre, and Kamila Palys.; writing—review and editing, Jenna Vesey, Tina Nosrati, Tiffany LaTorre, and Kamila Payls .; visualization, Kamila Palys, Tina Nosati, and Jenna Vesey.; All authors have read and agreed to the published version of the manuscript.”

**Data Availability Statement:**

The dataset can be downloaded as a CSV file under the name “StressLevelDataset.csv” at the link: https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis/data

**References**

1. Students experiencing stress. (n.d.). Https://Www.Apa.Org. Retrieved December 18, 2023, from https://www.apa.org/ed/schools/primer/stress
2. How stress affects your body and behavior. (n.d.). Mayo Clinic. Retrieved December 18, 2023, from https://www.mayoclinic.org/healthy-lifestyle/stress-management/in-depth/stress-symptoms/art-20050987
3. Studies show normal children today report more anxiety than child psychiatric patients in the 1950’s. (n.d.). Https://Www.Apa.Org. Retrieved December 18, 2023, from https://www.apa.org/news/press/releases/2000/12/anxiety
4. Student mental health is in crisis. Campuses are rethinking their approach. (n.d.). Https://Www.Apa.Org. Retrieved December 18, 2023, from https://www.apa.org/monitor/2022/10/mental-health-campus-care
5. Acharya, Chhabi. “Student Stress Factors: A Comprehensive Analysis.” *Kaggle*, 14 Oct. 2023, www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis/data.
6. Alharthy, Nesrin, et al. “Assessment of Anxiety Level of Emergency Health-Care Workers by Generalized Anxiety Disorder-7 Tool.” International Journal of Applied & Basic Medical Research, vol. 7, no. 3, July 2017, pp. 150–54. EBSCOhost, https://doi.org/10.4103/2229-516X.212963.
7. Zhang, Xijuan, et al. “Comparing the Psychometric Properties of a Scale Across Three Likert and Three Alternative Formats: An Application to the Rosenberg Self-Esteem Scale.” Educational & Psychological Measurement, vol. 83, no. 4, Aug. 2023, pp. 649–83. EBSCOhost, https://doi.org/10.1177/00131644221111402.
8. Wang, Larry, et al. “Screening for Perinatal Depression with the Patient Health Questionnaire Depression Scale (PHQ-9): A Systematic Review and Meta-Analysis.” General Hospital Psychiatry, vol. 68, Jan. 2021, pp. 74–82. EBSCOhost, https://doi.org/10.1016/j.genhosppsych.2020.12.007.
9. Esler, Murray, et al. “Chronic Mental Stress Is a Cause of Essential Hypertension: Presence of Biological Markers of Stress.” Clinical & Experimental Pharmacology & Physiology, vol. 35, no. 4, Apr. 2008, pp. 498–502. EBSCOhost, https://doi-org.library2.ramapo.edu:2443/10.1111/j.1440-1681.2008.04904.x.