**Analysis and Prediction of Stress Level Based on Students' Stress Factors**

Group Members: Kamila Palys, Tina Nosrati, Jenna Vesey, Tiffany LaTorre

**Introduction:**

In this project, our primary objective is to build a supervised classification predictive model aimed at predicting students' stress levels. With stress levels of students being rapidly on the rise, understanding how stress levels fluctuate and what are the primary drivers of this increased stress can prove crucial to future students' mental well-being. By comparing various factors, our model seeks to identify students’ risk of mental health issues and potentially determine how likely the student is in need of stress management support. Beyond predicting mental health concerns, our broader goal is to assess and quantify the importance of each contributing factor to students' mental well-being. Understanding the significance of individual variables in determining stress levels is important for preventing the emergence of future mental health issues as well as developing novel ways to treat this increase in stress. This comprehensive dataset not only enhances our predictive capabilities but also provides valuable insights for developing strategies to support students' mental health and well-being.

**Source and Background of Data:**

The Student Stress Factors dataset, integral to our project, is accessible on Kaggle via the following link: [Student Stress Factors – A Comprehensive Analysis](https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis/data). This data consists of 1100 data points, 20 features, and a target variable for stress level. This data is measured on a categorical scale, most being out of 30, in which students rated various factors based on their levels of stress related to each category. The closer the value is to 30, the more stressful this factor is to the student, while 0 indicates not stressful whatsoever. The target variable, stress levels, is ranked on a scale of 0-2. This information makes the dataset a valuable resource for our supervised classification predictive model. Each variable is categorized as a Social, Academic, Environmental, or Physiological Factor, offering a comprehensive perspective on the diverse elements contributing to stress. This dataset originates from a nationwide survey within Nepal encompassing 1100 students, ensuring its relevance to the broader student population. These students were mostly high school and college students. Most of the college students' data was obtained from Tribhuvan University. Our hope is that this dataset could be used to provide insight into the stress level of students from a wide range of countries, not just Nepal. This dataset will enable us to draw meaningful insights and construct a robust model for predicting and understanding student stress factors.

**Problems to Address:**

Our group believes in the need for a consistent predictive model capable of tracking the influential factors affecting students' stress levels. By developing such a model, we aim to pinpoint student factors that are most vulnerable to increased levels of stress, potentially help alleviate rising mental health issues, and tailor effective approaches to address students’ specific needs. A comprehensive understanding of stress-contributing factors can pave the way for the development of new techniques and strategies within educational settings for limiting stress levels. There are various factors that go into stress, and although we recognize this data set is not all encompassing for all the factors that go into stress, our aim is to at least identify the main factors of stress. Ultimately, our project seeks to build an understanding of the main issues within stress and to hopefully lead to the improvement of support systems for students, promoting their mental well-being and creating a healthier learning environment.

**Scientific Questions:**

1. What are the most critical factors contributing to students' stress levels?
2. How do these factors impact students' stress levels?
   1. *How do psychological factors like anxiety level, self-esteem, and depression contribute to predicting overall student stress levels?*
   2. *Are there specific patterns or thresholds in physiological data that correlate strongly with high-stress levels?*
   3. *How do social factors like social support, peer pressure, and extracurricular activities impact student stress levels?*
   4. *Are there particular aspects of the environment that have a more significant impact on stress?*
   5. *How accurately can academic performance, study load, and teacher-student relationships predict student stress levels?*
3. Are social, academic, environmental, psychological, or physiological factors most responsible for contributing to a student’s stress level?
4. Is there any correlation between stress factors in different aspects of the variables, such as social, academic, environmental, psychological, or physiological factors?
   1. *How do combinations of factors from different categories (psychological, physiological, social, environmental, and academic) collectively influence student stress?*
5. How can we predict the stress level in students?

**Techniques and Process:**

1. **Problem Definition [Jenna]**
   * This section will provide background information on understanding why managing stress is important. Stress is an average part of daily life, but recent studies have shown that stress is ticking. Stress leads to a lot of mental health conditions as well as is believed to exacerbate physical conditions. Although stress is something that needs to be examined, it is a complex psychological problem. To try to come up with novel treatments and stress management tactics, a better understanding of what factors contribute to stress and to what extent. Findings from this research will hopefully be able to help scientists develop ways to manage stress for students.
2. **Data Exploration** 
   * In this section of the project, we will use visualization and statistical techniques to find underlying patterns within the data. In addition to the relationship between variables, we will also search for outliers. Our final report will include significant findings of this step.

In data exploration, we will start with checking the dataset's basics, like its size, types of data, and if there are any null values within the dataset. This will help us understand the overall structure and direct us on how to begin processing this data. We will use numerical and visual summaries like the five-number summary and bar charts to visualize and explore each feature's key characteristics since our variables are categorical. This will help identify patterns as well as outliers and null values within the dataset itself.

Handling missing data responsibly and addressing outliers will be part of the process. We will also undergo data preparation techniques in order to ensure the dataset is clean and prepared for predictive modeling for any assumptions required for our specific modeling methods. Each group member will focus on a specific category of variables: social [Jenna], psychological [Jenna], environmental [Kamila], academic [Tiffany], and physiological factors[Tiffany].

1. **Machine Learning Algorithms** 
   * First, we will use unsupervised methods on the dataset to identify any initial patterns. This will help us learn more about our data and identify what variables may play a more significant role in stress levels than others. This will be helpful in comparison with our supervised classification techniques. We will perform clustering and study the attributes of these clusters, and finally, consider our target variable to see if observations within a cluster actually have the same stress level.
     1. K-modes Clustering [Tina]
   * Since our problem is a supervised classification problem, the machine-learning techniques we apply will include techniques including, but not limited to:
     1. Decision Tree [Tiffany]
     2. Random Forest [Kamila]
     3. KNN [Jenna]
     4. SVM [Tina] #assumptions not true
2. **Performance Improvement** 
   * In this section, we will focus on improving the performance of our predictive models using techniques including:
     1. Cross-validation [Tiffany]
     2. Ensemble Learning [Tina]

Cross-validation techniques will ensure that our models can effectively generalize to new and unseen data. Cross-validation could also be used to justify if our data is applicable to the entire country of Nepal for higher education students versus just the students that were surveyed in the data. We will use Ensemble methods to combine multiple predictors using voting mechanisms. This strategy aims to not only improve the accuracy of our models but also enhance their overall robustness and reliability in predicting student stress levels. The specific ensemble method will be determined upon our results, at which point in time, we will determine if we are going to use hard or soft voting to check the performance of our predicting models.

1. **Deployment and Testing [Each member for their specific machine learning technique]**
   * Lastly, each team member will test their classifier on test cases that were put aside from the dataset. In addition, we will create an applied function that accepts new input information from the user and can predict their stress level as an interactive piece from all of the information we got from our data. This way, potentially more data can be acquired, further testing the limits of our model.
2. **Final Report** 
   * Finally, we will conclude our project by preparing documentation and a written final report. Each member of the group will write about the tasks they performed.