**Final project – Students Mental Health Assessments**

Name: Dimple Krishna Chintala

Student ID: 4221938

Date: 12-13-2023

**Executive Summary**

The purpose of this report is to present findings on how or on what criteria they are chosen for the position, based on data that was made available for study. The characteristics that have the most effects on the desired column level are identified using a logistic regression model.

**Introduction:**

The dataset used in this project comprises 6995 rows and 20 columns, capturing various attributes that potentially impact an individual's mental health and well-being. Mental health assessment and prediction are critical areas within healthcare and psychology, influencing overall quality of life. This dataset aims to explore the relationship between different factors and mental health outcomes.

**Dataset Overview:**

The dataset encompasses a diverse range of attributes, each contributing uniquely to understanding mental health-related factors. It includes subjective perceptions (e.g., stress level, sleep quality), lifestyle elements (e.g., physical activity, diet quality), familial influences (e.g., family history), and external stressors (e.g., financial stress). Understanding these factors can aid in predictive modeling and mental health analysis.

**Attribute Breakdown:**

**Subjective Perception Attributes:**

Stress\_Level, Depression\_Score, Anxiety\_Score: Represent individual perceived stress, depression, and anxiety levels, vital in assessing mental health status and emotional well-being.

Sleep\_Quality: Reflects the quality of sleep, playing a significant role in mental and physical health.

**Lifestyle and Behavioral Attributes:**

Physical\_Activity, Diet\_Quality: Indicate the individual's physical fitness level and dietary habits, impacting overall health, including mental well-being.

Substance\_Use: Provides insights into substance consumption patterns affecting mental health outcomes.

**Social and Environmental Factors:**

Social\_Support, Family\_History: Represent the support system and genetic predisposition, affecting mental health resilience and vulnerabilities.

Chronic\_Illness, Financial\_Stress: Reflects chronic health conditions and financial strain, influencing mental health and stress levels.

**Academic Attribute:**

Semester\_Credit\_Load: This represents academic workload, which could influence stress levels and mental well-being during academic sessions.

**Utilization in the Project:**

These attributes serve as crucial features for predictive modeling, helping build machine learning models to predict mental health outcomes. Leveraging these features can enhance the accuracy of predictions and provide valuable insights into factors influencing mental health.

**Relevance to Mental Health Prediction:**

Understanding and analyzing these attributes aids in uncovering patterns and correlations between lifestyle, environmental factors, and mental health. Predictive modeling utilizing these attributes could offer valuable tools for early detection, intervention, and personalized mental health care.

**Distribution visualization for numerical variables:**

**1. Age Distribution:**

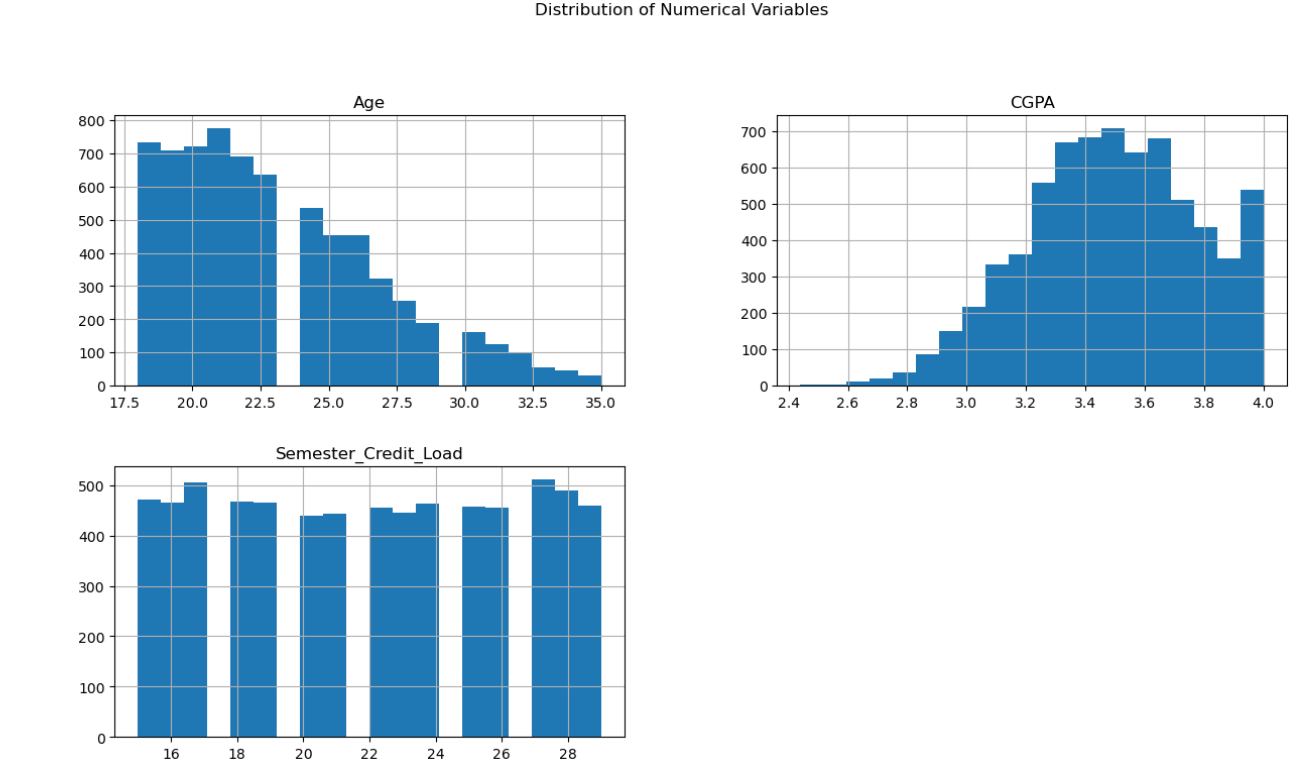
The histogram for 'Age' represents the distribution of ages among the individuals in your dataset. It shows how many individuals fall into specific age ranges or bins. The x-axis typically represents different age groups or values, and the y-axis represents the frequency or count of individuals within each age group.

**2. CGPA Distribution:**

This histogram portrays the distribution of Cumulative Grade Point Average (CGPA) among the individuals. It visualizes how the CGPA values are distributed across the dataset. Each bar in the histogram represents a range of CGPA values, and the height of the bar indicates the number of individuals within that CGPA range.

**3. Semester Credit Load Distribution:**

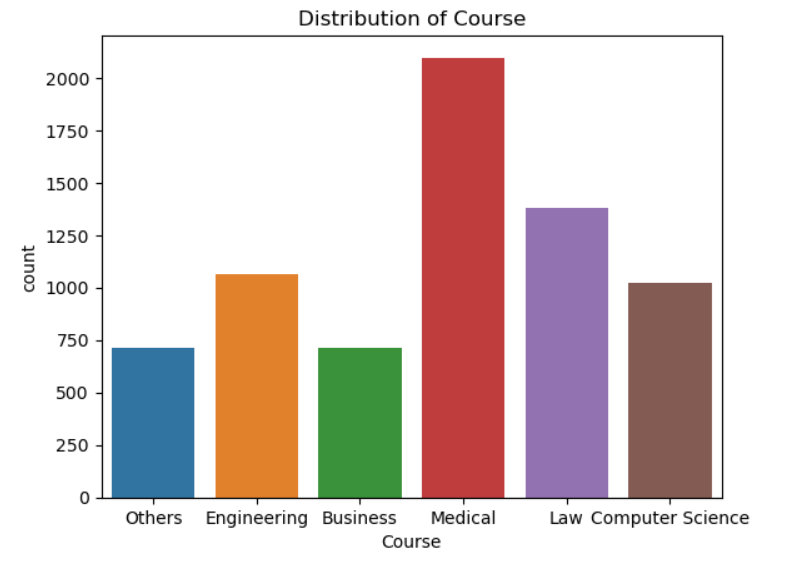
This histogram showcases the distribution of the number of semester credits students are enrolled in. It displays how many individuals are taking particular credit loads, with bins representing different credit load ranges. The histogram helps visualize the concentration or dispersion of students across various credit loads.

****

**Bar plots for categorical variables**

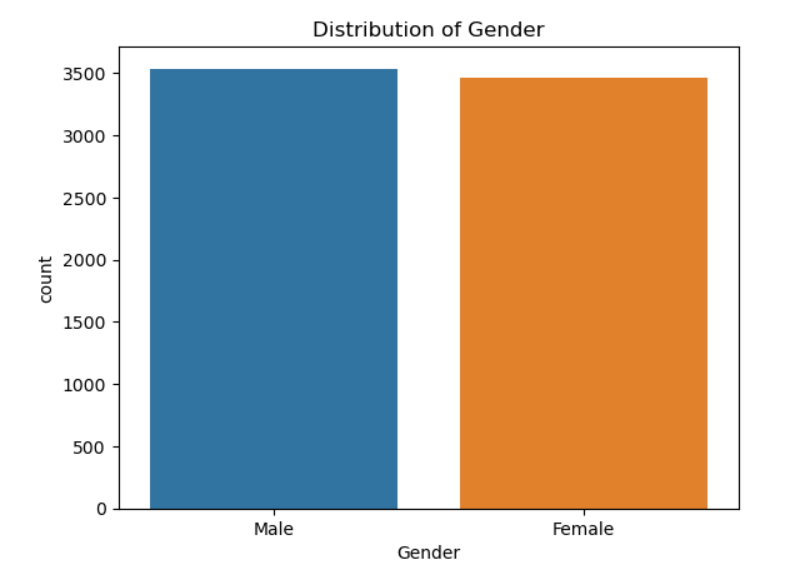
**1. Course:**

The bar plot for 'Course' illustrates the count or frequency of students enrolled in different academic courses or majors. Each bar corresponds to a specific course, showing how many individuals are pursuing that particular academic program.

****

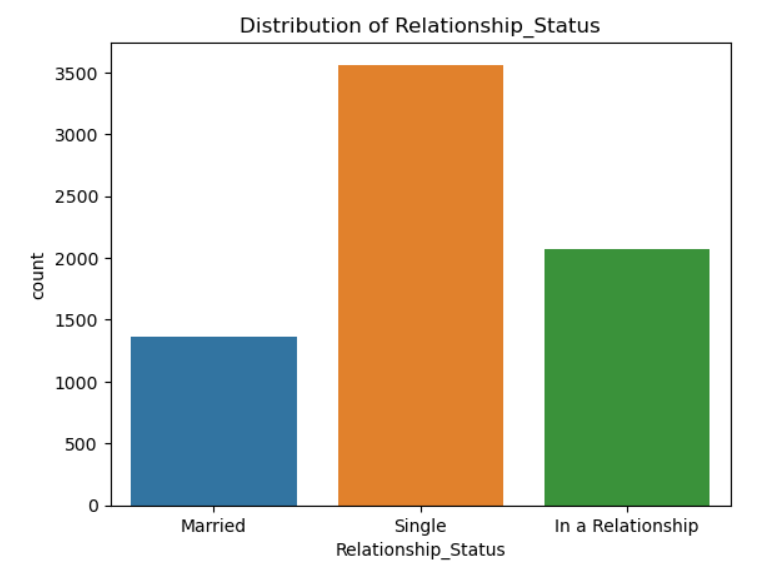
**2. Gender:**

This bar plot showcases the distribution of genders within the dataset. It represents the count or frequency of individuals categorized by gender, displaying the number of males and females or other gender identities.



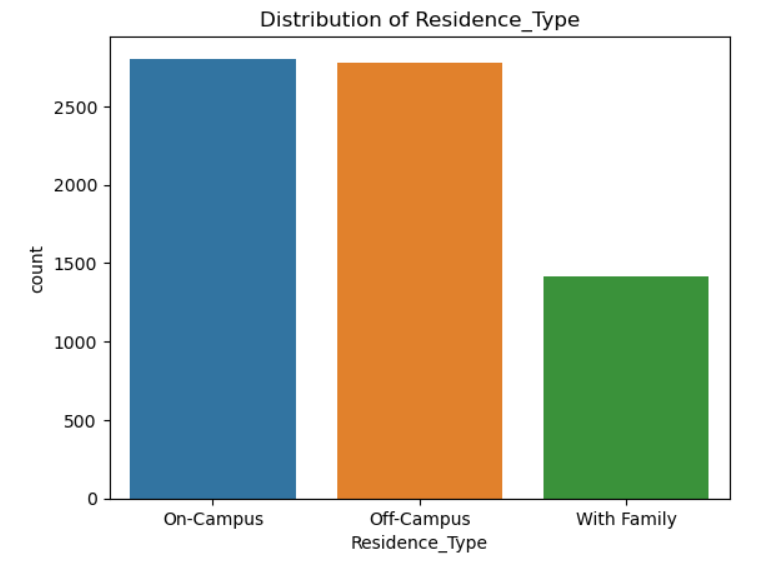
**3. Relationship\_Status:**

The bar plot for 'Relationship\_Status' represents the distribution of individuals based on their relationship status. It shows the count or frequency of individuals categorized by whether they are single, in a relationship, married, or other relationship statuses.



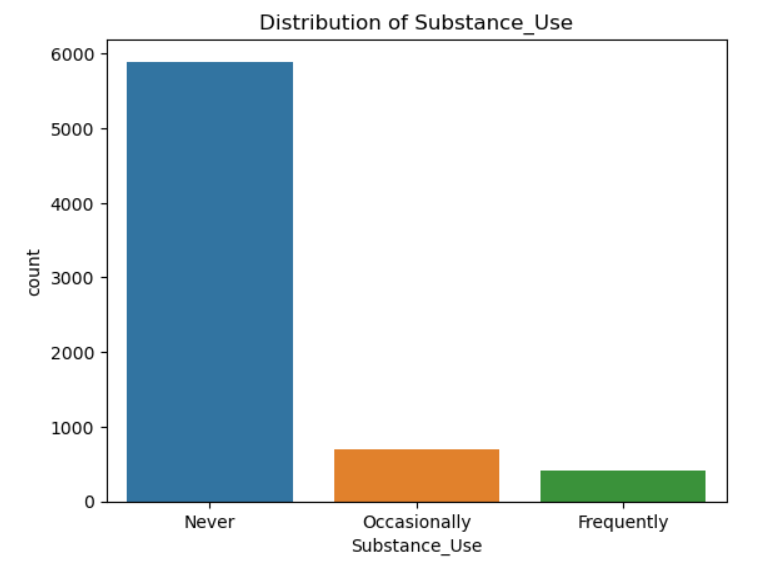
4. Residence\_Type:

This bar plot illustrates the count or frequency of individuals based on their residence type. It could display categories such as 'On-Campus', 'Off-Campus', or 'With Family', showing where individuals reside during their academic term.



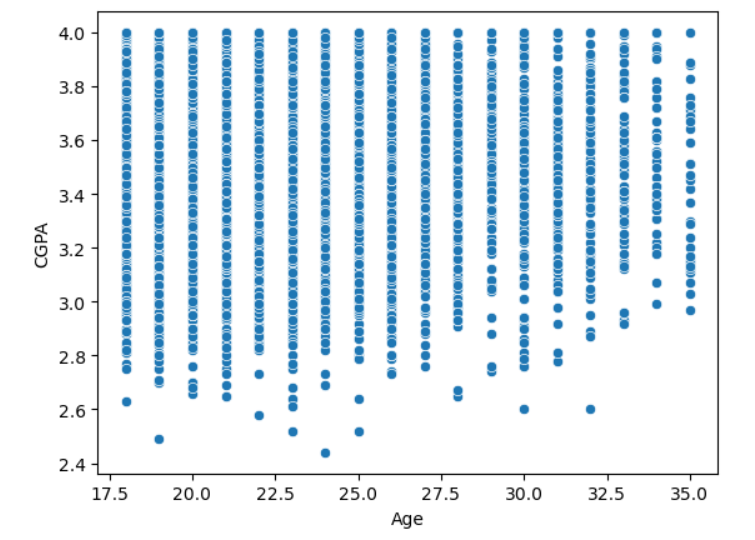
5. Substance\_Use:

The bar plot for 'Substance\_Use' represents the count or frequency of individuals based on their reported frequency or extent of substance use. It might show categories such as 'Never', 'Occasionally', or 'Frequently', indicating the usage patterns among the individuals in the dataset.



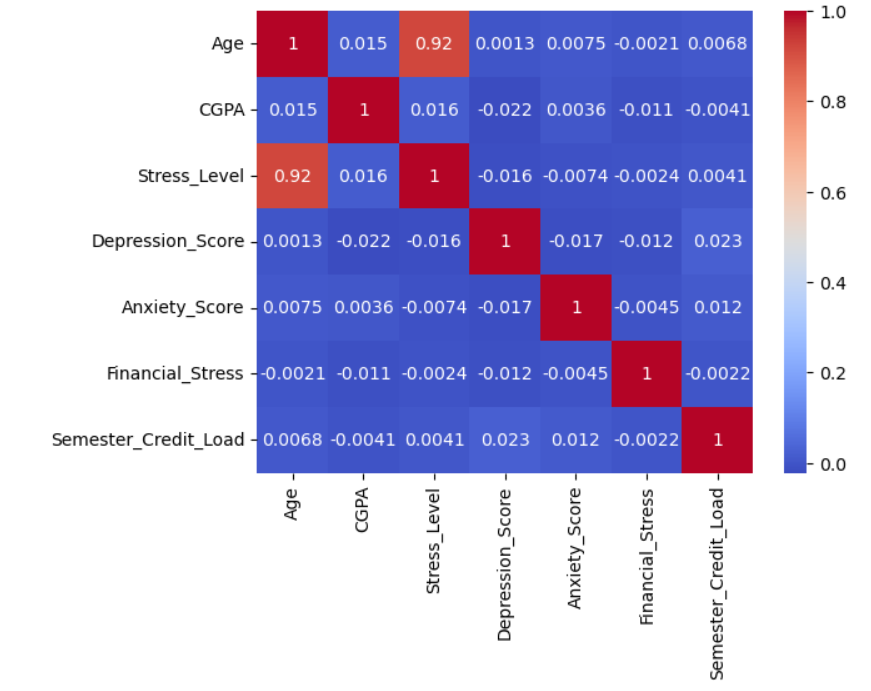
**Scatter plot for Age vs CGPA:**

* Data Points: Each point on the scatter plot represents a student's data with their age on the x-axis and CGPA on the y-axis.
* Trend Observation: By examining the distribution of points, look for any discernible patterns or trends. A typical trend might show higher CGPA scores among certain age groups or vice versa.
* Correlation: Assess whether there's a visible correlation between age and academic performance (CGPA). A positive correlation might suggest that as age increases, CGPA tends to increase as well, or it might reveal no apparent relationship.



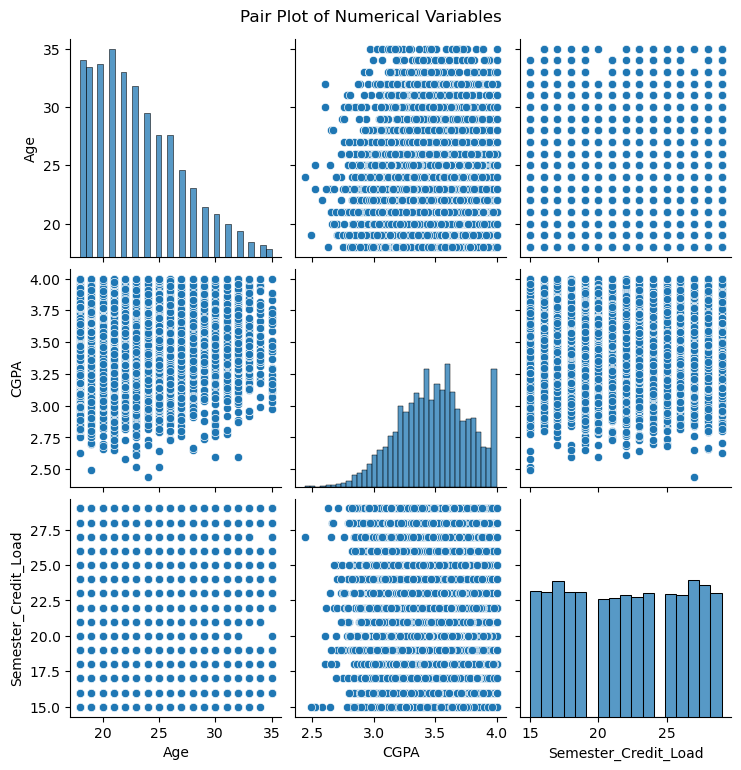
**Correlation matrix for numeric columns:**

* Positive Correlation (close to +1): Variables move in the same direction. As one increases, the other tends to increase as well.
* Negative Correlation (close to -1): Variables move in opposite directions. As one increases, the other tends to decrease.
* No Correlation (close to 0): Variables show no linear relationship.



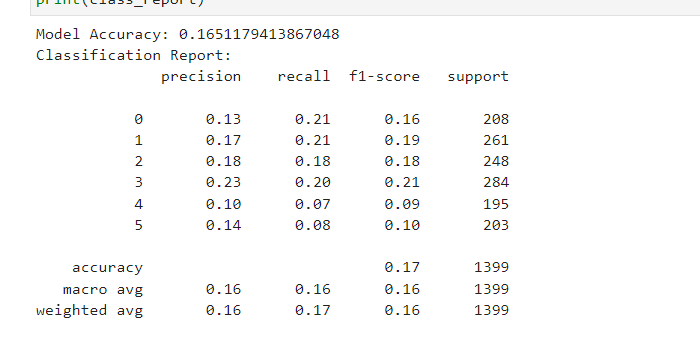
* Strongly correlated variables are age with stress level it is positively correlated with 92% .

**Pair plot numerical variables explanation :**



**Model fitting:**

KNN model:



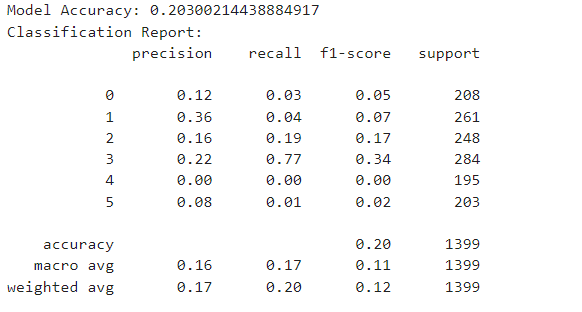
The K-Nearest Neighbors (KNN) model you've trained seems to have an accuracy of around 16.5%. Looking at the classification report:

* Precision: Reflects the accuracy of the model in predicting each class. For instance, for class 0, about 13% of the instances predicted as class 0 were correct.
* Recall: Indicates the proportion of actual instances of a class that were predicted correctly. For example, around 21% of actual instances of class 0 were correctly predicted.
* F1-score: Represents the harmonic mean of precision and recall. Higher F1-scores denote better balance between precision and recall.
* Support: The number of instances of each class in the test set.

Both the 'macro avg' and 'weighted avg' metrics give an overall summary across all classes. In this case, the model's performance across different classes seems quite similar, with precision, recall, and F1-scores ranging between approximately 0.1 to 0.23.

The accuracy and other metrics suggest that the model might be having challenges distinguishing between different stress levels based on the provided features. Improving this could involve trying different algorithms, feature engineering, scaling or transforming features, or optimizing the model's hyperparameters to better capture the relationships between the features and stress levels.

**Logistic regression model:**



* Accuracy: Around 20.3%, slightly higher than previous models.
* Precision: Varies for different classes. For instance, class 3 has a precision of approximately 22%, class 1 has a precision of around 36%, while classes 0, 2, 4, and 5 have lower precision scores.
* Recall: Varies significantly across classes, with class 3 having the highest recall of about 77%, class 2 at 19%, and the others relatively low.
* F1-score: Also varies across classes, with class 3 having a relatively higher F1-score of approximately 34%, while other classes have lower scores.
* Support: Indicates the number of instances for each class in the test set.

The 'macro avg' and 'weighted avg' metrics give an overall summary across all classes, demonstrating that the model performs relatively better on class 3 compared to the other classes, given its higher recall and F1-score.

The overall performance of this logistic regression model appears better than the KNN models previously discussed. However, there is still room for improvement, especially in classes where the precision, recall, and F1-scores are low. Techniques like feature engineering, hyperparameter tuning, handling class imbalance (if present), or trying different models altogether could potentially enhance the model's performance.

**Conclusion:**

If the project's primary aim is to predict stress levels, the conclusion could emphasize the focal point of stress level prediction:

The main goal of this study was to forecast stress levels using a variety of characteristics. The study that went into analyzing these variables revealed how complex it is to predict stress levels using the datasets that are currently available. The difficulty of precisely forecasting different levels of stress was brought to light by the models created for stress level prediction, such as logistic regression and KNN.

While logistic regression was more accurate, it had trouble accurately predicting the stress level for several categories. While KNN performed more evenly across a range of stress levels, it struggled to discern subtle differences between them.

The results highlight the complex interplay between several variables and stress, indicating the complex network influencing stress levels. These findings highlight the continuous need to improve stress level forecast accuracy by utilizing a variety of analytical approaches, researching complex feature engineering techniques, and adjusting models.

Even though the models' ability to forecast outcomes was limited, the knowledge they provided was important in understanding stress dynamics. This knowledge might help create more potent stress-reduction plans by providing deeper insights into stress prediction and promoting mental health."