# Predictive Data Analysis

Predictive Analytics refers to the field that applies various quantitative methods on data to make real-time predictions.

2023-10-30

.gdbar img{ width: 170px !important; height: 100px !important; margin: 4px 4px; }
.gdbar{ width: 200px !important; height: 100px !important; }

# Introduction to the Dataset

## Exploring Student Stress

This dataset dives into the factors affecting the stress levels of students. It comprises approximately 20 essential features carefully chosen from various aspects of students' lives, scientifically categorized into Psychological, Physiological, Social, Environmental, and Academic factors.

## Key Aspects of the Dataset:

Psychological Factors: Understanding mental well-being through features like 'anxiety_level,' 'self_esteem,' 'mental_health_history,' and 'depression.'

Physiological Factors: Exploring physical health indicators including 'headache,' 'blood_pressure,' 'sleep_quality,' and 'breathing_problem.'

Environmental Factors: Analyzing the impact of surroundings with variables such as 'noise_level,' 'living_conditions,' 'safety,' and 'basic_needs.'

Academic Factors: Determining academic pressures and concerns using 'academic_performance,' 'study_load,' 'teacher_student_relationship,' and 'future_career_concerns.'

Social Factors: Understanding peer dynamics and social support through 'social_support,' 'peer_pressure,' 'extracurricular_activities,' and 'bullying.'

# Structure

In this dataset, we explore various aspects of students' lives to understand what causes stress. By using predictive analysis, we want to find patterns and make predictions. Our goal is to help create a better and supportive environment for students.

```
## 'data.frame':    1100 obs. of  21 variables:
##  $ anxiety_level               : int  14 15 12 16 16 20 4 17 13 6 ...
##  $ self_esteem                 : int  20 8 18 12 28 13 26 3 22 8 ...
##  $ mental_health_history       : int  0 1 1 1 0 1 0 1 1 0 ...
##  $ depression                  : int  11 15 14 15 7 21 6 22 12 27 ...
##  $ headache                    : int  2 5 2 4 2 3 1 4 3 4 ...
##  $ blood_pressure              : int  1 3 1 3 3 3 2 3 1 3 ...
##  $ sleep_quality               : int  2 1 2 1 5 1 4 1 2 1 ...
##  $ breathing_problem           : int  4 4 2 3 1 4 1 5 4 2 ...
##  $ noise_level                 : int  2 3 2 4 3 3 1 3 3 0 ...
##  $ living_conditions           : int  3 1 2 2 2 2 4 1 3 5 ...
##  $ safety                      : int  3 2 3 2 4 2 4 1 3 2 ...
##  $ basic_needs                 : int  2 2 2 2 3 1 4 1 3 2 ...
##  $ academic_performance        : int  3 1 2 2 4 2 5 1 3 2 ...
##  $ study_load                  : int  2 4 3 4 3 5 1 3 3 2 ...
##  $ teacher_student_relationship: int  3 1 3 1 1 2 4 2 2 1 ...
##  $ future_career_concerns      : int  3 5 2 4 2 5 1 4 3 5 ...
##  $ social_support              : int  2 1 2 1 1 1 3 1 3 1 ...
##  $ peer_pressure               : int  3 4 3 4 5 4 2 4 3 5 ...
##  $ extracurricular_activities  : int  3 5 2 4 0 4 2 4 2 3 ...
##  $ bullying                    : int  2 5 2 5 5 5 1 5 2 4 ...
##  $ stress_level                : int  1 2 1 2 1 2 0 2 1 1 ...
```

# Psychological Factors (Column Description)

**anxiety_level:** Measures the level of anxiety a student experiences, ranging from 0 (low anxiety) to 21 (high anxiety).(HADS-A Score)

**self_esteem:** Indicates the level of self-esteem of the student, ranging from 0 (low self-esteem) to 30 (high self-esteem).

**mental_health_history:** Binary indicator (0 or 1) representing whether the student has a history of mental health issues.

**depression:** Measures the degree of depression the student is facing, ranging from 0 (low depression) to 27 (high depression).

# Physiological Factors (Column Description)

**headache:** Frequency of headaches experienced by the student, ranging from 0 (no headaches) to 5 (frequent headaches).

**blood_pressure:** Blood pressure level of the student, with values ranging from 1 (low) to 3 (high).

**sleep_quality:** Rates the quality of the student's sleep on a scale from 0 (poor quality) to 5 (excellent quality).

**breathing_problem:** Indicates whether the student faces breathing problems, with values 0 (no) or 1 (yes).

# Environmental Factors (Column Description)

**noise_level:** Perception of noise levels in the student's environment, ranging from 0 (low noise) to 5 (high noise).

**living_conditions:** Assessment of the student's living conditions, with values from 0 (poor conditions) to 5 (excellent conditions).

**safety:** Perceived safety level of the student's surroundings, ranging from 0 (unsafe) to 5 (very safe).

**basic_needs:** Satisfaction level of the student's basic needs, from 0 (not satisfied) to 5 (fully satisfied).

# Academic Factors (Column Description)

**academic_performance:** Student's self-perceived academic performance, with values from 0 (poor) to 5 (excellent).

**study_load:** Student's perception of their study workload, ranging from 0 (light) to 5 (heavy).

**teacher_student_relationship:** Quality of relationship with teachers, with values from 0 (poor) to 5 (excellent).

**future_career_concerns:** Concerns about future career prospects, ranging from 0 (low concern) to 5 (high concern).

# Social Factors (Column Description)

**social_support:** Level of social support experienced by the student, from 0 (low support) to 3 (high support).

**peer_pressure:** Influence of peer pressure on the student, with values from 0 (low pressure) to 5 (high pressure).

**extracurricular_activities:** Student's participation in extracurricular activities, ranging from 0 (no participation) to 5 (active participation).

**bullying:** Experience of bullying by the student, with values from 0 (no bullying) to 5 (frequent bullying).

**stress_level:** Overall stress level reported by the student, with values from 0 (low stress) to 2 (high stress).

# Objective

Determine which specific factors have the most significant impact on students' stress levels. This insight is vital for understanding the primary drivers of stress among students.

## Problem Statement:

Identify and prioritize key factors influencing students' stress levels to develop targeted interventions and support systems within educational institutions.

## STEP 1 - Data Cleaning

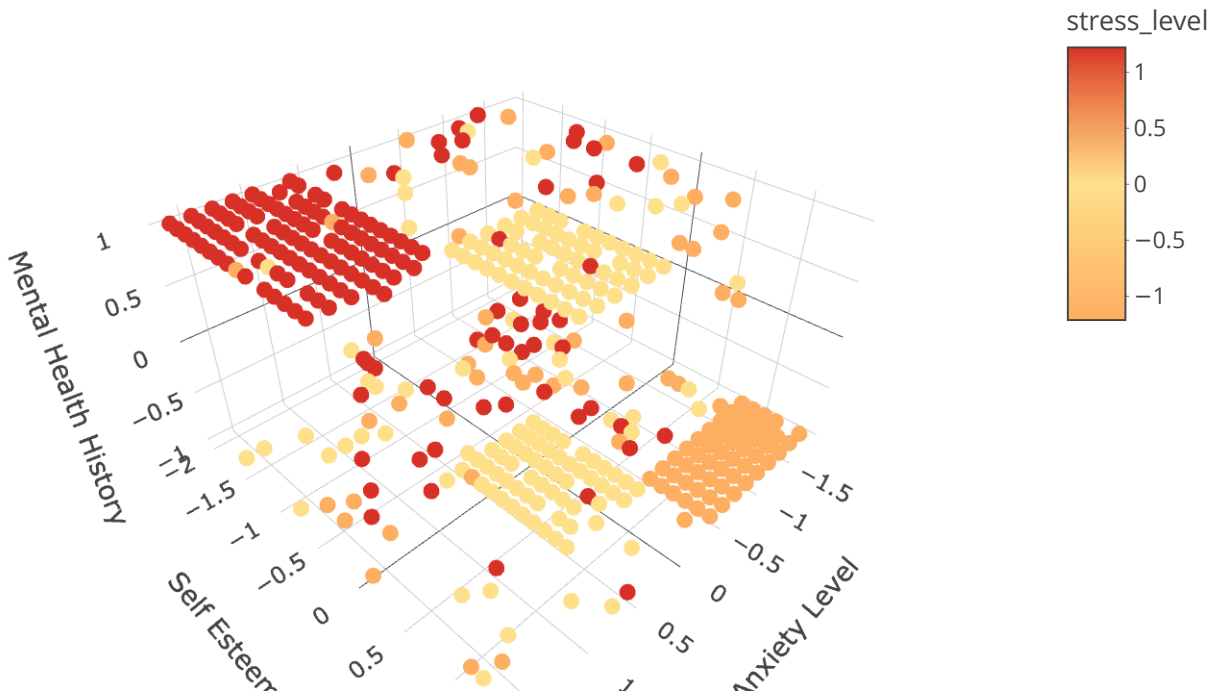
-First Step would be to clean the data but the data is already clean as it has no null values or outliers.

## STEP 2 - Data Preparation

-Second Step would be standardization i.e. standardizing the dataset to follow the same format for better visualization and training of our predictive model.

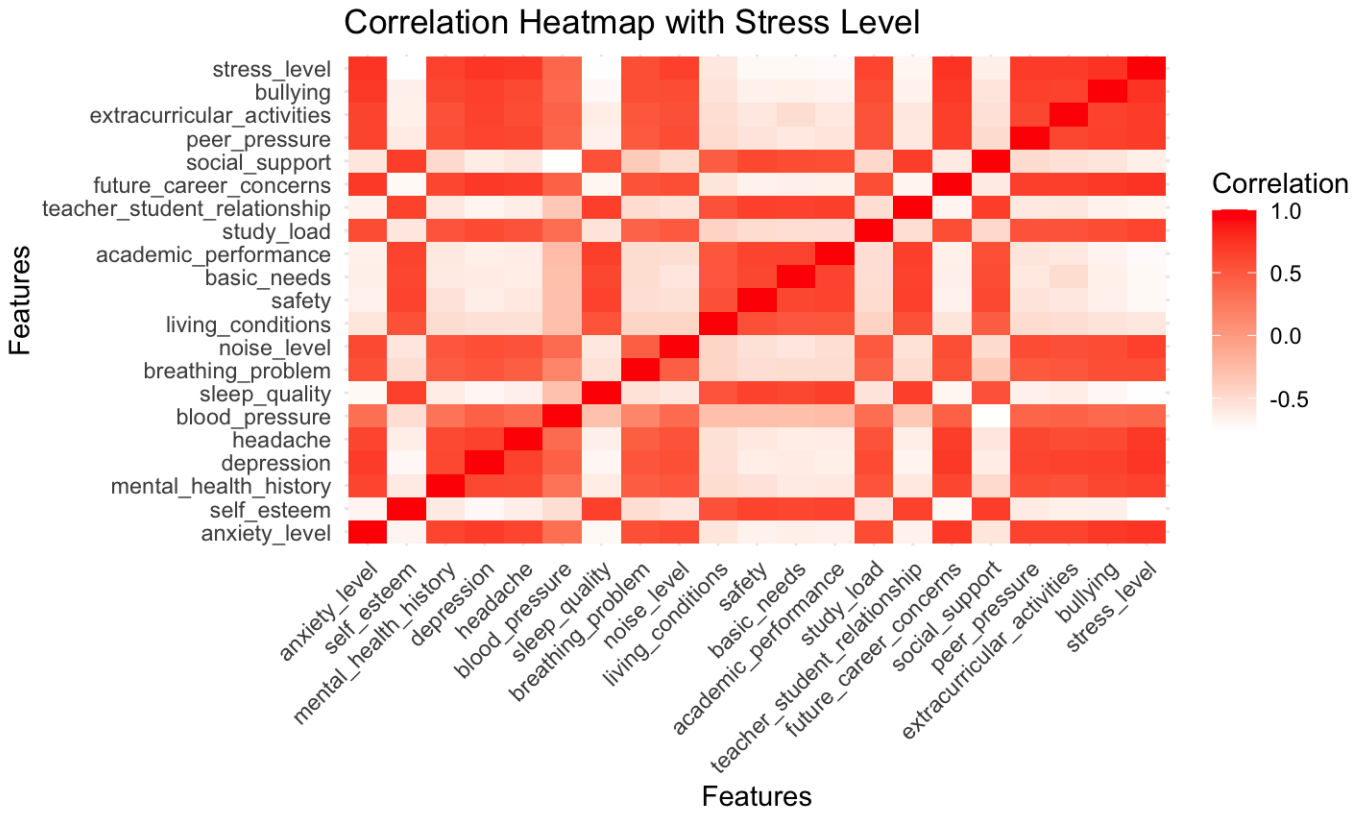
The scale() function centers and scales the columns of a dataset, making them have a mean of 0 and a standard deviation of 1. Now our data is ready for predictive analysis.

# STEP 3 - Data Exploration (Part 1)




The 3D scatter plot allows you to visualize the relationship between three variables: anxiety_level, self_esteem, and stress_level. Clusters indicate groups of students with similar anxiety levels, self-esteem, and stress levels.

# STEP 3 - Data Exploration (Part 2)



# Insights

The correlation coefficient between stress level and health indicators is positive and significant for most of the health indicators. This means that as stress level increases, so do the health indicators. This suggests that stress is associated with a number of negative health consequences.

The strongest correlation is between stress level and burnout, with a correlation coefficient of 0.675. This suggests that burnout is a very common consequence of stress.

Other health indicators that are strongly correlated with stress level include:

-Depression (correlation coefficient = 0.606)
-Anxiety (correlation coefficient = 0.598)
-Headaches (correlation coefficient = 0.485)
-Sleep quality (correlation coefficient = -0.479)

The negative correlation between stress level and sleep quality suggests that stress can lead to poor sleep quality. Poor sleep quality, in turn, can lead to increased stress levels. This can create a vicious cycle.

# STEP 4 and STEP 5

## STEP 4 - Data Splitting

The training set (train_data) contains 80% of the data, and the testing set (test_data) contains the remaining 20%.

## STEP5 - Model Selection and Building (Desicion Tree)

```
## n= 881
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 881 601.8865000 0.98864930
##    2) self_esteem>=24.5 300  45.3966700 0.13666670
##      4) sleep_quality>=3.5 262   7.8625950 0.02290076 *
##      5) sleep_quality< 3.5 38  10.7631600 0.92105260 *
##    3) self_esteem< 24.5 581 226.2857000 1.42857100
##      6) bullying< 3.5 305  65.9475400 1.01311500 *
##      7) bullying>=3.5 276  49.5181200 1.88768100
##       14) basic_needs>=3.5 7   0.8571429 0.14285710 *
##       15) basic_needs< 3.5 269  26.7955400 1.93308600
##         30) living_conditions>=2.5 16  10.0000000 1.00000000 *
##         31) living_conditions< 2.5 253   1.9841900 1.99209500 *
```

# STEP 6 Model Evaluation (Regression)

```
mse <- mean((test_data$stress_level - predictions)^2)
print(paste("Mean Squared Error: ", mse))
```

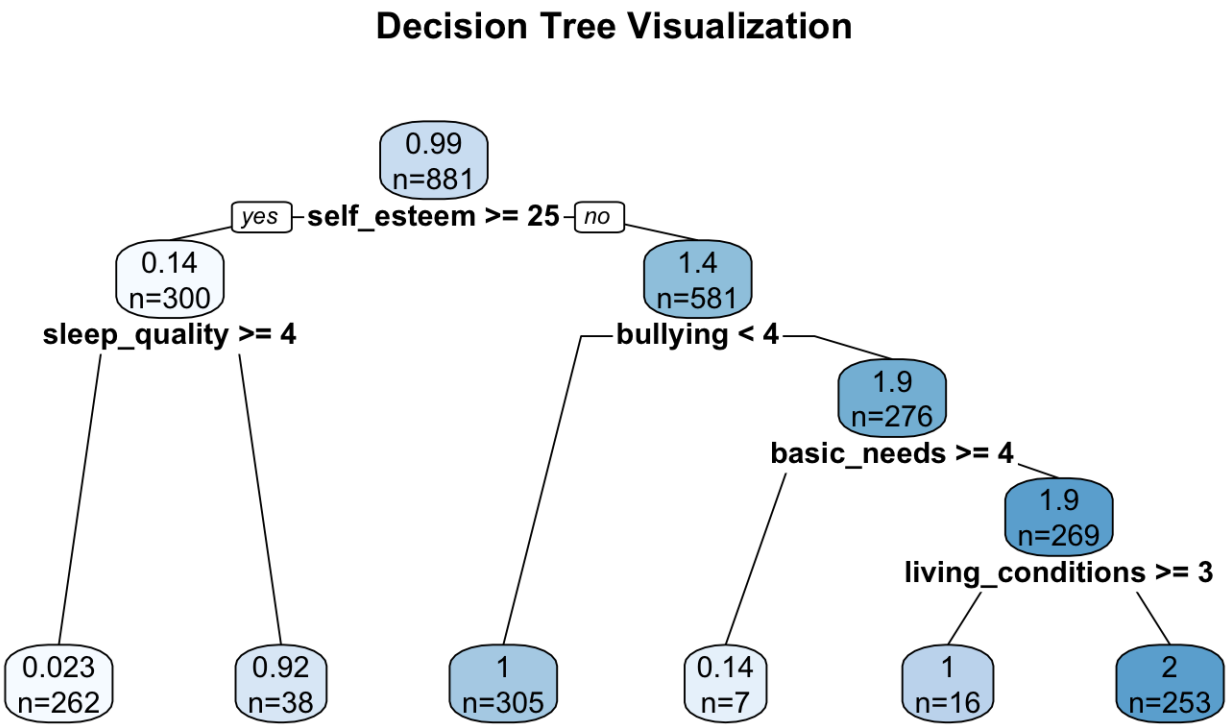
```
## [1] "Mean Squared Error:  0.194221062913345"
```

The MSE value represents the average squared difference between the predicted stress levels and the actual stress levels in the test data.

A lower MSE indicates that, on average, the model's predictions are closer to the actual stress levels.

In this case, an MSE of 0.196 suggests that, on average, the squared difference between the predicted and actual stress levels is quite small.

# STEP 7 - Visualization 1

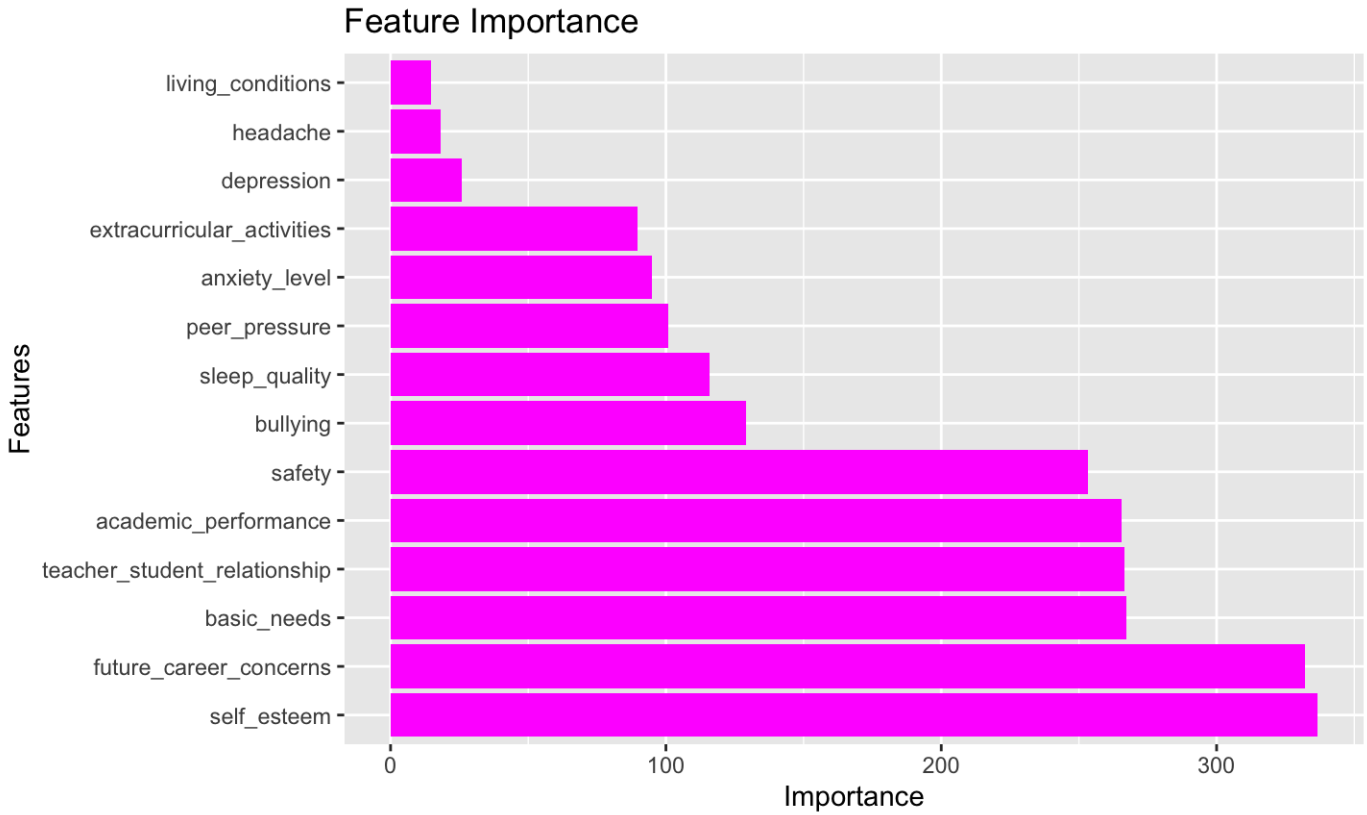


# STEP 7 - Interpretation 1

A few interpretation that can be made from the descision tree:

- Students who are involved in extracurricular activities are predicted to have lower stress levels.
- Students who are not involved in extracurricular activities are predicted to have higher stress levels.
- Students with high academic performance are predicted to have lower stress levels.
- Students with low academic performance are predicted to have higher stress levels.
- Students with high self-esteem are predicted to have lower stress levels.
- Students with low self-esteem are predicted to have higher stress levels.
- Students with met basic needs are predicted to have lower stress levels.
- Students with unmet basic needs are predicted to have higher stress levels.
- Students with good living conditions are predicted to have lower stress levels.
- Students with poor living conditions are predicted to have higher stress levels.
- Students who are bullied are predicted to have higher stress levels.

# STEP 7 - Visualization 2




**Interpretation** The plot shows that the two most important features in predicting stress levels are living conditions and future career concern. This is consistent with the interpretation of the decision tree itself.

# Learning and Findings

In this predictive data analysis project, we delved into a comprehensive dataset exploring multiple facets of students' lives, aiming to unravel the intricate web of factors contributing to their stress levels. Through exploratory data analysis, correlation analysis, and a decision tree regression model, we gained valuable insights into the influential aspects of student stress.

## Key Findings:

Extracurricular Activities & Academic Performance: Engagement in extracurricular activities correlates with lower stress levels. Conversely, students with lower academic performance tend to experience higher stress.

Psychological Factors: Factors such as self-esteem play a crucial role. Students with higher self-esteem exhibit lower stress levels.

Basic Needs & Living Conditions: Meeting basic needs and living in favorable conditions are associated with reduced stress.

Social Dynamics: Bullying negatively impacts stress levels, indicating the importance of addressing social pressures.

Future Concerns: Worries related to future career prospects elevate stress levels among students.

# Useful Recommendations/Insights:

Interventions and Support: Schools should focus on promoting extracurricular involvement, enhancing self-esteem, and providing resources to meet students' basic needs.

Mental Health Support: Schools should offer counseling services to address psychological factors and cope with academic stressors.

Bullying Prevention: Implement anti-bullying programs and create a supportive environment to reduce social pressures.

Career Counseling: Provide comprehensive career guidance and counseling to alleviate students' concerns about their future.

Policy Implementation: Institutions should consider these findings while formulating policies aimed at reducing student stress, fostering a healthier learning environment.

# References

https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis/data