Anxiety Prediction Of Student Behaviour Based On

Day To Day Life

Sanskruti Mirajkar   
 *Electronics and Communication*  
*KLE Technological University*: Belgaum,India 02FE22BEC084@kletech.ac.in

Shraman Kanthi  
*Electronics and Communication*   
*KLE Technological University*Belgaum,India  
02FE22BEC092@kletech.ac.in

Darshan Modekar  
*Electronics and Communication*   
*KLE Technological University*Belgaum,India  
02FE22BEC119@kletech.ac.in Swati Patil  
*Electronics and Communication*   
*KLE Technological University*Belgaum,India  
02FE22BEC112@kletech.ac.in

*Abstract*— The prevalence of anxiety levels among student-ts has significantly increased in recent years, emphasizing the need for timely detection and intervention. This study explores the prediction of anxiety levels in students using various machine learning algorithms, including K-Nearest Neighbors (K-NN), Naive Bayes and Logistic Regression. The dataset, collected through targeted questionnaires, encompasses factors such as Cumulative Grade Point Average (CGPA), age, and other demographic variables, alongside symptoms of anxiety. By transforming these attributes into quantifiable parameters, the models were trained and will be evaluated for their effectiveness in predicting anxiety. Among the algorithms, the best one with more high accuracy and less error and best performance will be selected.

This project underscores the potential of machine learning techniques in identifying students at risk of anxiety based on their lifestyle and academic performance indicators. By providing actionable insights, this study aims to contribute to the development of effective mental health support strategies. The importance of this project lies in its capacity to facilitate early intervention, helping students manage anxiety more effectively and promoting overall well-being and academic success.

Keywords— anxiety, K-NN, Logistic Regression, Naive Bayes.

# Introduction

In today’s fast-paced world, the mental health of individuals, especially students, is increasingly at risk. Academic pressures, career ambitions, and personal relationship issues are significant contributors to anxiety, depression, and stress. For students, these stressors are often magnified due to the challenges of balancing daily academic responsibilities, future career planning, and social expectations. According to the World Health Organization, depression is among the most prevalent mental health disorders, affecting over 300 million people globally, and is often accompanied by anxiety and stress. These mental health conditions can hinder a student’s ability to focus, learn, and maintain a healthy lifestyle, impacting overall academic performance and well-being.

Numerous predictive models have been developed to address mental health disorders, particularly focusing on common symptoms such as feelings of guilt, worthlessness, helplessness, restlessness, and even suicidal thoughts. Generalized Anxiety Disorder (GAD), for instance, manifests through symptoms like nervousness, sleeplessness, gastrointestinal issues, and lack of concentration. Stress, similarly, is characterized by symptoms such as inability to relax, irritability, fatigue, chronic headaches, and frequent infections. Differentiating these disorders accurately poses a challenge for machines, as many symptoms overlap among anxiety, depression, and stress.

To address this, we propose a machine learning-based approach to predict anxiety levels in students by analyzing behavioral patterns in daily life. Utilizing the Depression, Anxiety, we aim to classify the severity of these mental health conditions into five categories. In this study, we apply machine learning algorithms such as Naïve Bayes, K-Nearest Neighbors (KNN) and Logistic Regression to predict and classify anxiety . The goal is to enhance accuracy in diagnosing these mental health issues among students and provide a foundation for effective intervention.

The paper is organized as follows: Section 2 discusses related research, Section 3 details the materials and methodology, Section 4 presents the results obtained from our predictive models, and Section 5 provides the conclusion. An accurate predictive model for mental health diagnosis is critical to better support student well-being in academic settings..

# literature survey

The rise in mental health issues, particularly anxiety, among students highlights the need for predictive models that facilitate early identification and intervention. Research demonstrates that machine learning models can effectively predict mental health conditions by analyzing behavioral and demographic data.

Reece et al. [6] used a Hidden Markov Model (HMM) to assess depression and PTSD among Twitter users, revealing the potential of social media data in identifying mental health risks. Braithwaite et al. [7] employed decision trees to evaluate suicide risk from tweets, achieving 92% accuracy, showing how text-based data can reveal mental health risks*.*

Sau et al. (2017) analyzed mental health among older adults using various models, finding random forest to be the most accurate at 91%, especially when handling complex datasets. Anu Priya et al. (2020) predicted anxiety, stress, and depression with algorithms such as KNN and random forest, selecting random forest for its effective handling of imbalanced data. Hou et al. [8] used reading habits to predict depression, with Naïve Bayes performing best, highlighting how specific behaviors can indicate mental health issues.

This study builds on these approaches by targeting student anxiety prediction through daily life factors, like CGPA and demographic variables. By assessing KNN, SVM, decision tree, Naïve Bayes, random forest, and logistic regression models, the study aims to identify the most effective model to support early interventions, promoting mental well-being and academic success*.*

# dataset description and preprocessing

In this paper, we used the dataset provided on Kaggle. This dataset is about the university students anxiety. The dataset has 1977 data(rows) and 39 features(columns).

|  |  |
| --- | --- |
| Column Name | Column Description |
| Age | Age of the participant. |
| Gender | Gender of the participant. |
| University | The name of the university where the individual is enrolled or graduated from |
| Department | The academic department or field of study the individual is part of, such as Psychology engineering etc. |
| Year | The current year of study or the academic year the individual is in |
| CGPA | Cumulative Grade Point Average |
| Scholarship | Indicates whether the individual is receiving a scholarship, typically a binary value (yes/no). |
| Nervous | A measure of the individual’s feelings of nervousness, potentially on a scale. |
| Worrying | Indicates the frequency or intensity of worrying thoughts experienced by the individual. |
| Relaxing | A measure of how often the individual engages in relaxing activities or feels relaxed. |
| Irritated | A measure of how often the individual feels irritated or annoyed. |
| Overthinking | Indicates the tendency of the individual to engage in overthinking or excessive rumination. |
| Restless | A measure of how often the individual feels restless or unable to relax. |
| Afraid | Indicates the frequency or intensity of feelings of fear experienced by the individual. |
| Anxiety Value | A numerical score quantifying the individual’s level of anxiety. |
| Anxiety label | A categorical label representing the individual’s anxiety level (e.g., low, moderate, high). |
| Upset | A measure of how often the individual feels upset or emotionally disturbed. |
| Control | Indicates the individual’s perceived level of control over their emotions or situation. |
| Stressed | A measure of how often the individual feels stressed or overwhelmed. |
| Cope | Indicates the individual’s ability to cope with stress or difficult situations. |
| Confident | A measure of the individual’s self-confidence or belief in their abilities. |
| Flow | Indicates the experience of being in a state of flow, characterized by deep engagement in activities. |
| Irritation Control | A measure of how well the individual can control feelings of irritation. |
| Performance | A measure of the individual’s performance in academic or other relevant contexts. |
| Anger | Indicates the frequency or intensity of feelings of anger experienced by the individual. |
| Difficulties | A measure of the difficulties the individual faces in their academic or personal life. |
| Stress Value | A numerical score quantifying the individual’s level of stress. |
| Stress Label | A categorical label representing the individual’s stress level (e.g., low, moderate, high). |
| Interest | A measure of the individual’s interest in their studies or activities. |
| Down | Indicates how often the individual feels down or low in mood. |
| Sleep | A measure of the quality or quantity of sleep the individual gets. |
| Energy | Indicates the individual’s energy levels, possibly affecting their daily activities. |
| Appetite | A measure of the individual’s appetite or eating patterns. |
| Self-Worth | Indicates the individual’s sense of self-worth or self-esteem. |
| Concentration | A measure of the individual’s ability to concentrate on tasks or studies. |
| Movement | Indicates the individual’s level of physical activity or restlessness. |
| Suicidal Thoughts | A measure of whether the individual experiences thoughts of self-harm or suicide. |
| Depression Value | A numerical score quantifying the individual’s level of depression. |
| Depression Label | A categorical label representing the individual’s depression level (e.g., low, moderate, high). |

Fig.: Table1

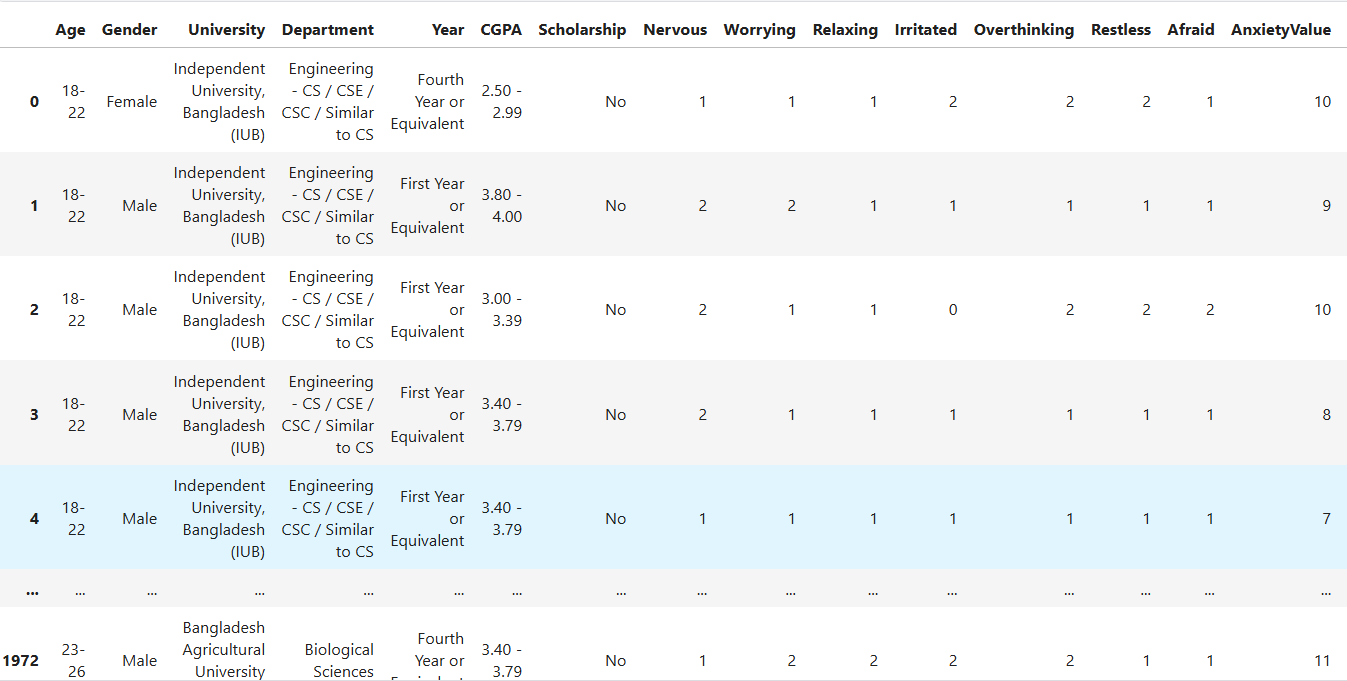


Fig.1. Dataset description

## Identifying duplicate values

Identifying duplicate values in a dataset is a crucial step in data preprocessing, especially in projects involving data analysis, machine learning, or statistical modeling. Duplicate values can arise due to data entry errors, system glitches, or merging datasets from multiple sources. These duplicates can introduce biases, inflate the significance of certain observations, and lead to misleading results in analysis. In the context of your project, which seems to involve analyzing academic performance and mental health attributes, duplicate entries could misrepresent patterns, such as the prevalence of anxiety or depression, or skew the relationship between variables like stress and academic performance. By identifying and removing duplicates, you ensure the dataset accurately represents unique individuals, leading to more reliable and valid insights. This step is particularly important when the goal is to draw meaningful conclusions about the population being studied or to build predictive models based on clean and unbiased data.

 Removing duplicate values is essential to ensure the dataset accurately represents unique observations, avoids bias, and maintains the integrity of the analysis or model. Therefore, we must delete duplicate values for reliable and valid results.

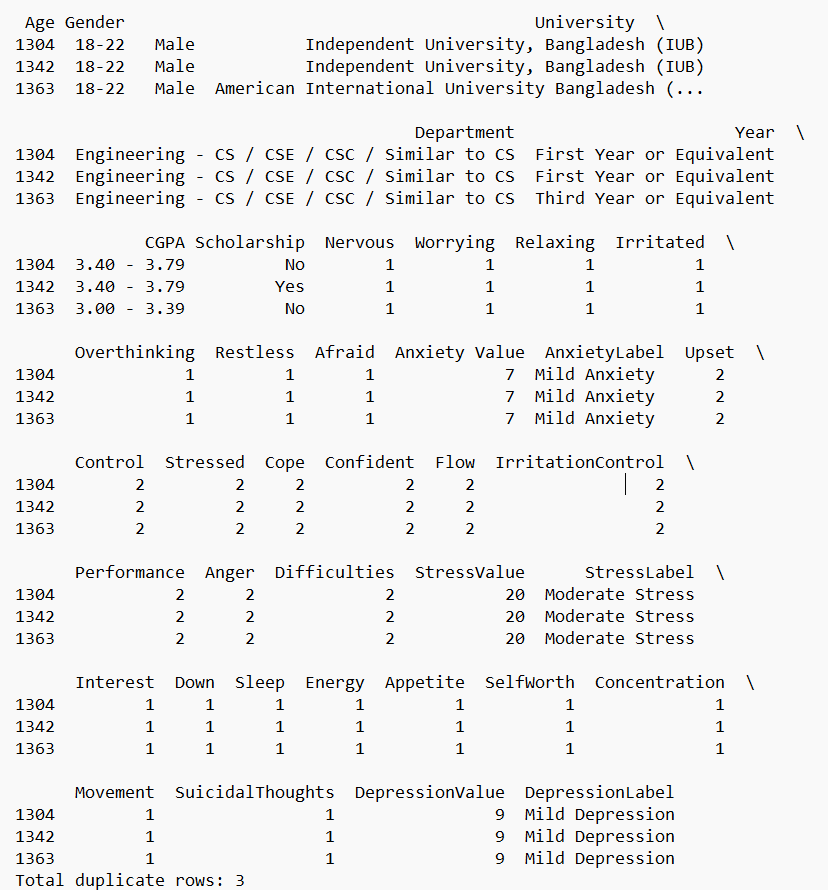


Fig.2. Duplicates values

## Identifying unique values

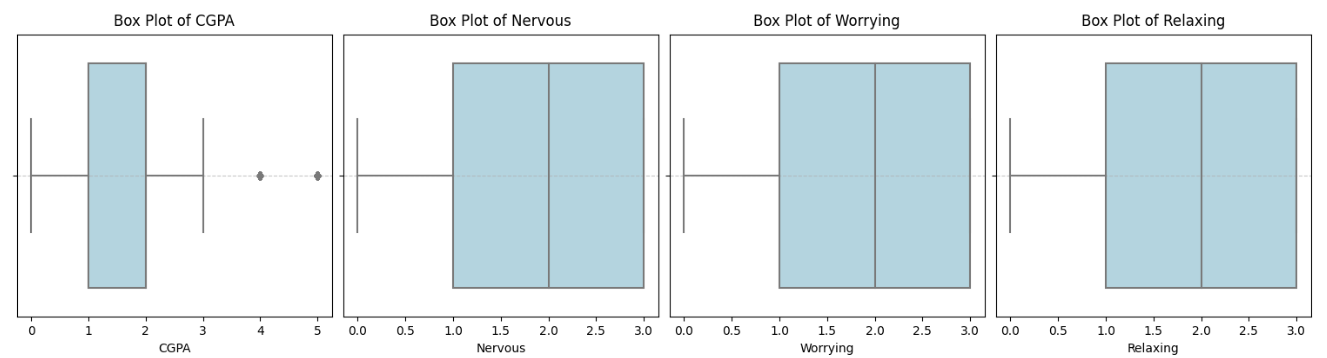
Identifying unique values in a dataset is vital for gaining meaningful insights and ensuring accurate analysis. In the context of our project, which focuses on academic performance and mental health attributes, unique values help us understand the diversity and range of responses, such as variations in anxiety, stress, or depression levels among individuals. This process allows us to identify patterns, correlations, and outliers while ensuring each data point is distinct and contributes to the overall findings. Recognizing unique values also prevents redundancy, supports better decision-making, and enhances the reliability of conclusions drawn from the data. Hence it gives us target column

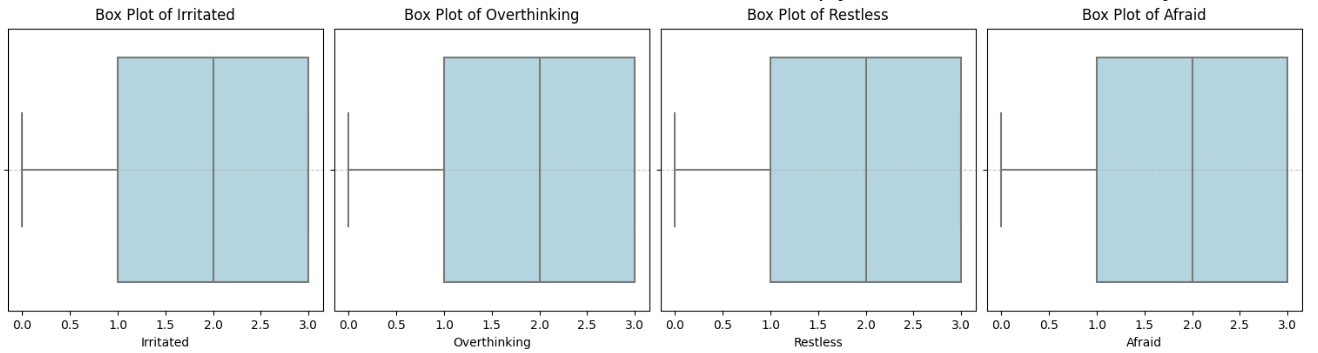
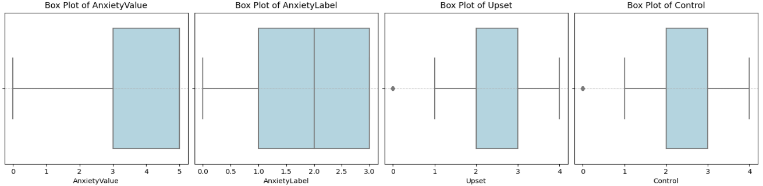
|  |  |
| --- | --- |
| Attributes | Unique Value |
| Age | 5 |
| Gender | 3 |
| University | 15 |
| Department | 12 |
| Year | 5 |
| CGPA | 6 |
| Scholarship | 2 |
| Nervous | 4 |
| Worrying | 4 |
| Relaxing | 4 |
| Irritated | 4 |
| Overthinking | 4 |
| Restless | 4 |
| Afraid | 4 |
| Anxiety Value | 22 |
| Anxiety label | 4 |
| Upset | 5 |
| Control | 5 |
| Stressed | 5 |
| Cope | 5 |
| Confident | 5 |
| Flow | 5 |
| Irritation Control | 5 |
| Performance | 5 |
| Anger | 5 |
| Difficulties | 5 |
| Stress Value | 38 |
| Stress Label | 3 |
| Interest | 4 |
| Down | 4 |
| Sleep | 4 |
| Energy | 4 |
| Appetite | 4 |
| Self-Worth | 4 |
| Concentration | 4 |
| Movement | 4 |
| Suicidal Thoughts | 4 |
| Depression Value | 28 |
| Depression Label | 6 |

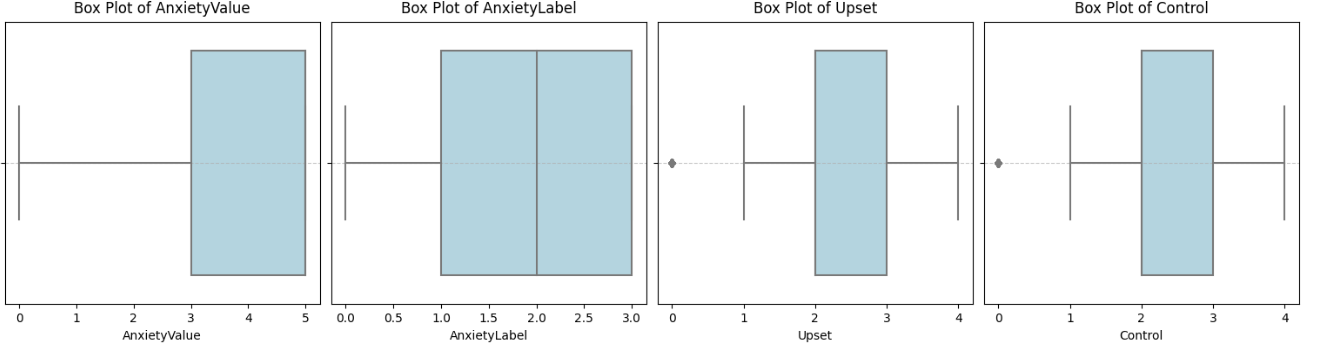
Fig.: Table2

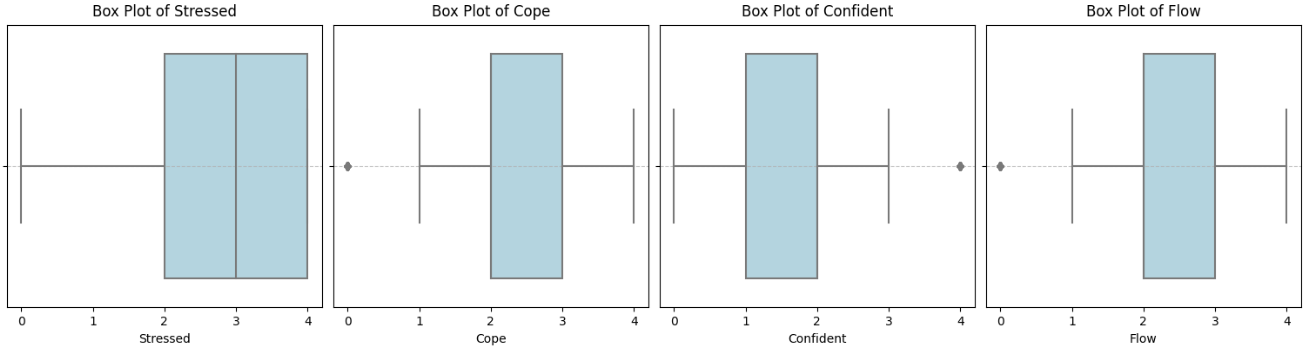
## Handling Outliers

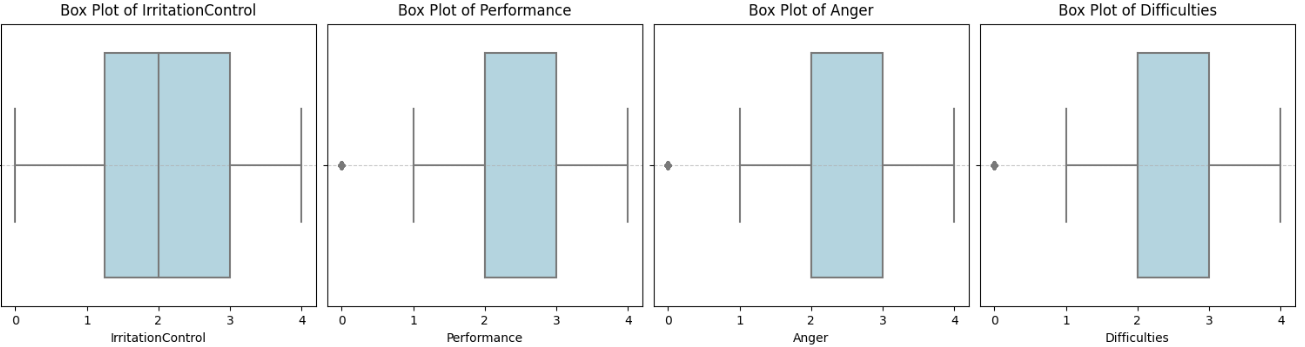
Handling outliers is a critical step in data preprocessing to ensure the accuracy and reliability of analysis. Outliers are extreme values that deviate significantly from the rest of the data and can arise due to measurement errors, data entry mistakes, or genuine variability. In our project, which examines academic performance and mental health attributes, outliers could distort statistical measures like mean and standard deviation, leading to misleading results. Proper handling, such as removing or transforming outliers based on their impact, ensures that the dataset accurately reflects underlying trends and relationships, improving the quality of insights and the performance of predictive models.

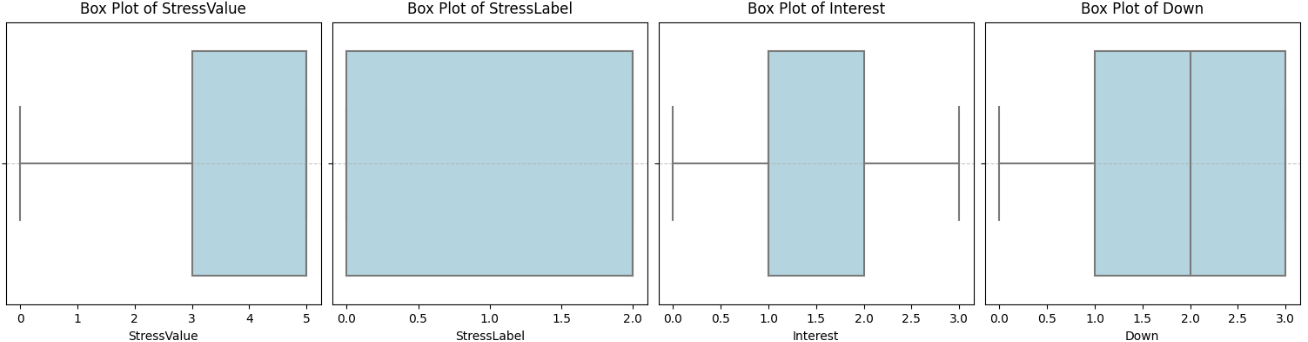


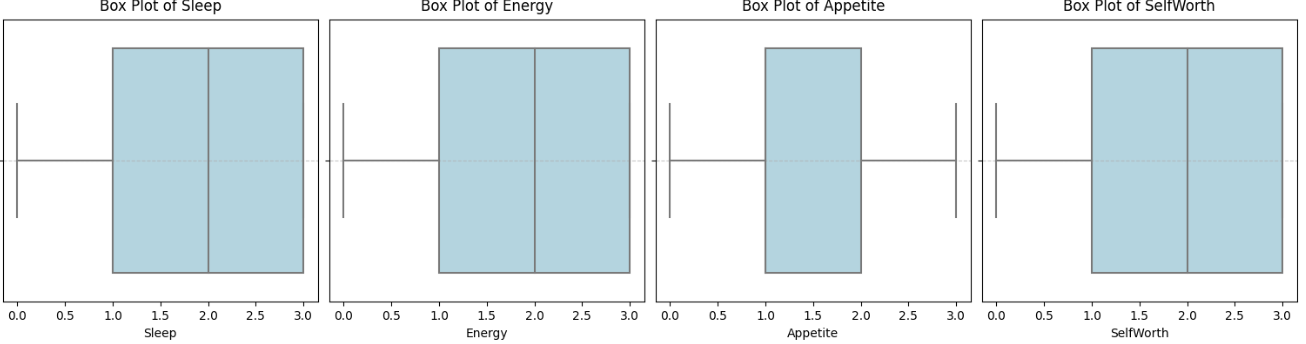
 











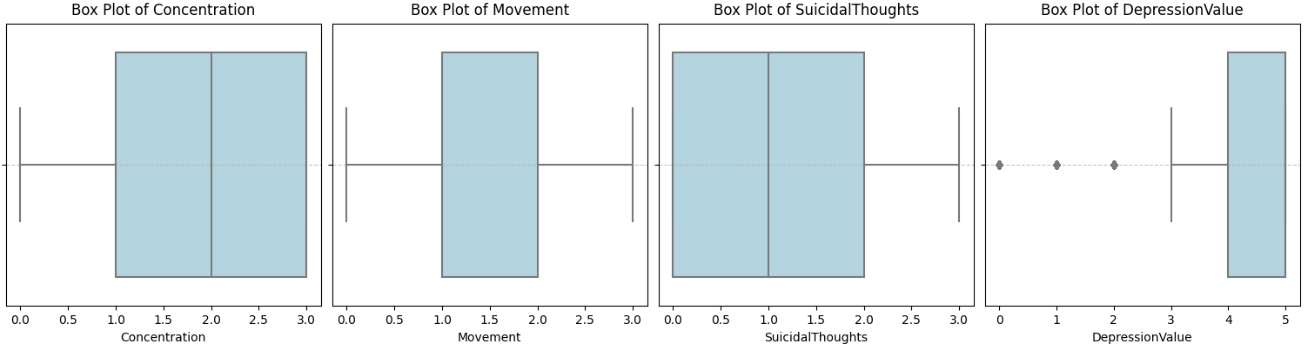


Fig.3. Box plot of outliers

Dropping outliers is essential to maintain the integrity of our dataset and enhance the accuracy of our analysis. Outliers can skew results, distort statistical calculations, and lead to misleading interpretations, particularly in our project focused on academic performance and mental health attributes. By removing these extreme values, we ensure that the remaining data more accurately represents the true patterns and relationships within the population, ultimately leading to more reliable insights and improved predictive modeling.

## Data Cleaning

Data cleaning is a fundamental process in data analysis that involves identifying and correcting errors or inconsistencies in a dataset to improve its quality and reliability. This process typically includes handling missing values, correcting typos, removing duplicates, and addressing outliers that may distort analysis results. In our project, which examines academic performance and mental health attributes, effective data cleaning ensures that the dataset accurately reflects the true characteristics of the population being studied. By standardizing formats, removing irrelevant information, and ensuring consistency across entries, we can enhance the dataset's usability, leading to more valid insights and better-informed decision-making. Ultimately, thorough data cleaning lays the groundwork for successful analysis, predictive modeling, and effective interpretation of results.

## Correlation Matrix

A correlation matrix is a table that displays the correlation coefficients between multiple variables in a dataset. Each cell in the matrix shows the correlation between two variables, with values ranging from -1 to +1. A correlation of +1 indicates a perfect positive correlation, meaning that as one variable increases, the other also increases. Conversely, a correlation of -1 signifies a perfect negative correlation, where one variable increases as the other decreases. A correlation of 0 indicates no relationship between the variables.

In our project, a correlation matrix is invaluable for understanding the relationships among various academic performance and mental health attributes, such as anxiety, stress, and depression levels. By visualizing these correlations, we can identify patterns and potential dependencies, helping to uncover insights about how these factors interact. This information is crucial for further analysis, hypothesis testing, and building predictive models, as it allows us to focus on the most relevant variables and better understand the dynamics at play in our data. Overall, a correlation matrix serves as a powerful tool for exploratory data analysis, guiding our research direction and informing subsequent steps in our project.

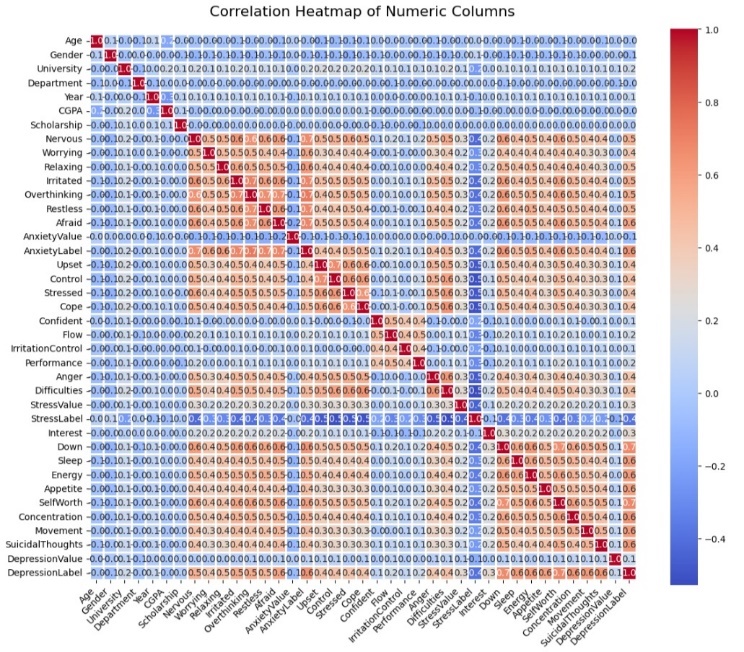


Fig.4. Heat map

From the above graph we came to a conclusion that there is a high correlation between the below attributes

Nervous, Worrying, Relaxing, Irritated, Overthinking, Restless, Afraid and Anxiety label, Upset, Control, Stressed, Cope, Confident, Down, Sleep, Energy, Appetite, Self-worth, Concentration, Movement, Depression Level. So out of all these attributes, I can retain single attributes so that it will be having summary of other attributes so that it can be correlated

# VISUALIZATION

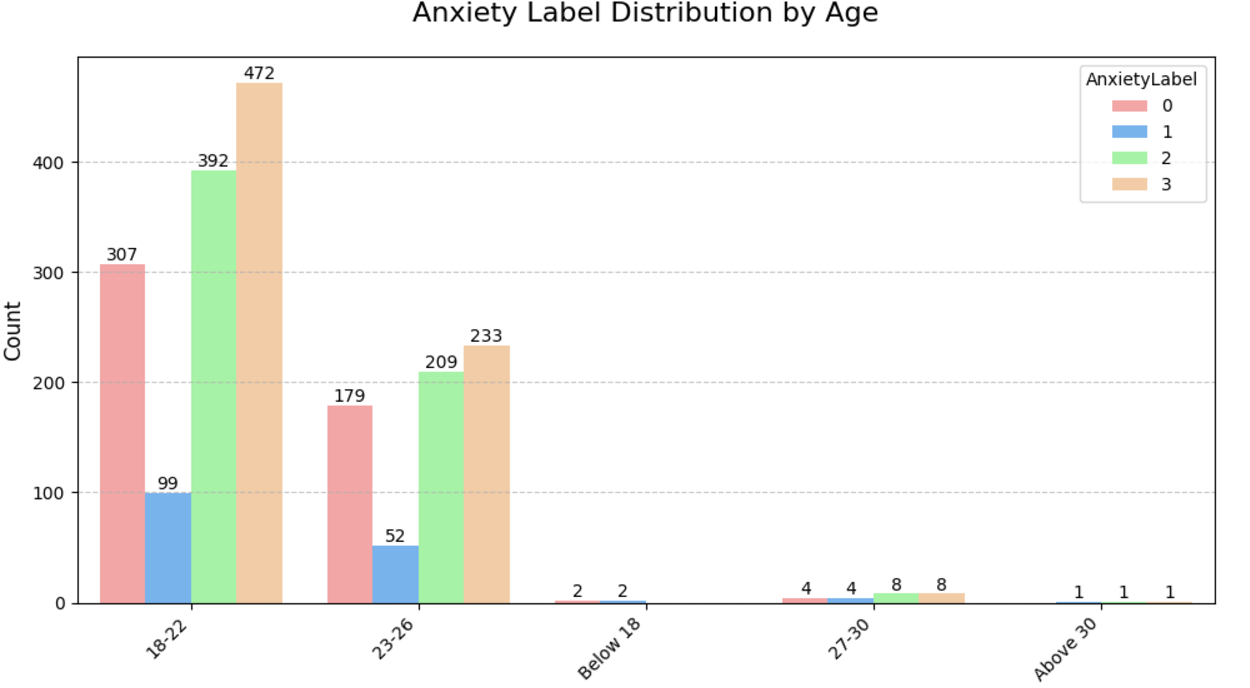


Fig.5. Anxiety label vs Age

The above Fig.5 shows the anxiety Vs Age distribution. Here we can identify that the age group of 18-22 have more students suffering from anxiety.

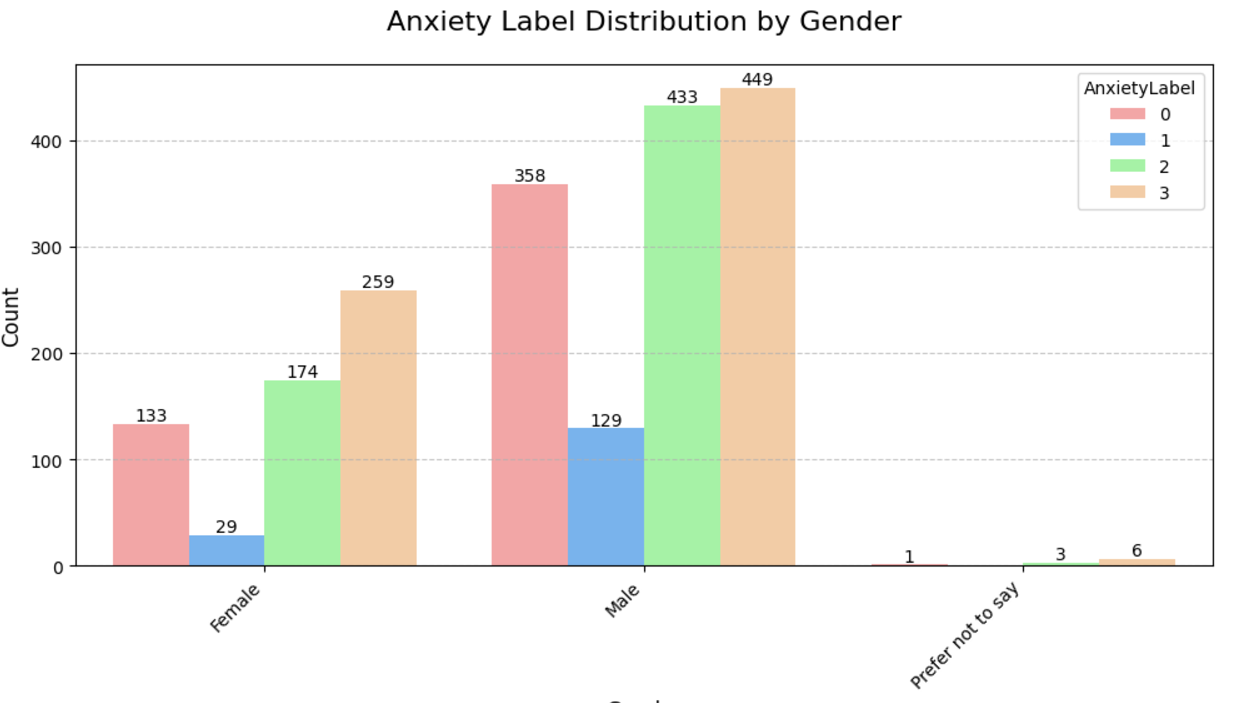


Fig.6. Anxiety label vs Gender

The above Fig.6 shows the anxiety Vs Gender distribution. Here we can identify that Male students suffer more from anxiety

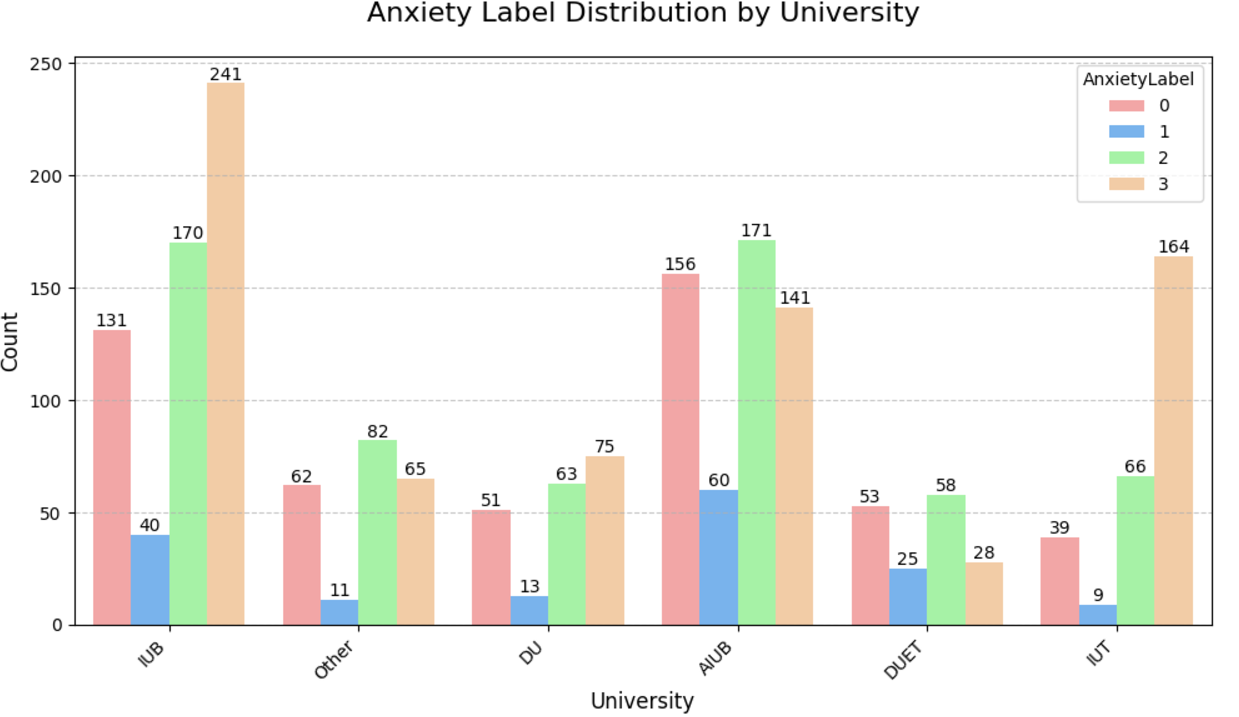


Fig.7. Anxiety label Vs University

The above Fig.7 show Anxiety label Vs University. Here we can identify that University of IUB is suffering more from Anxiety

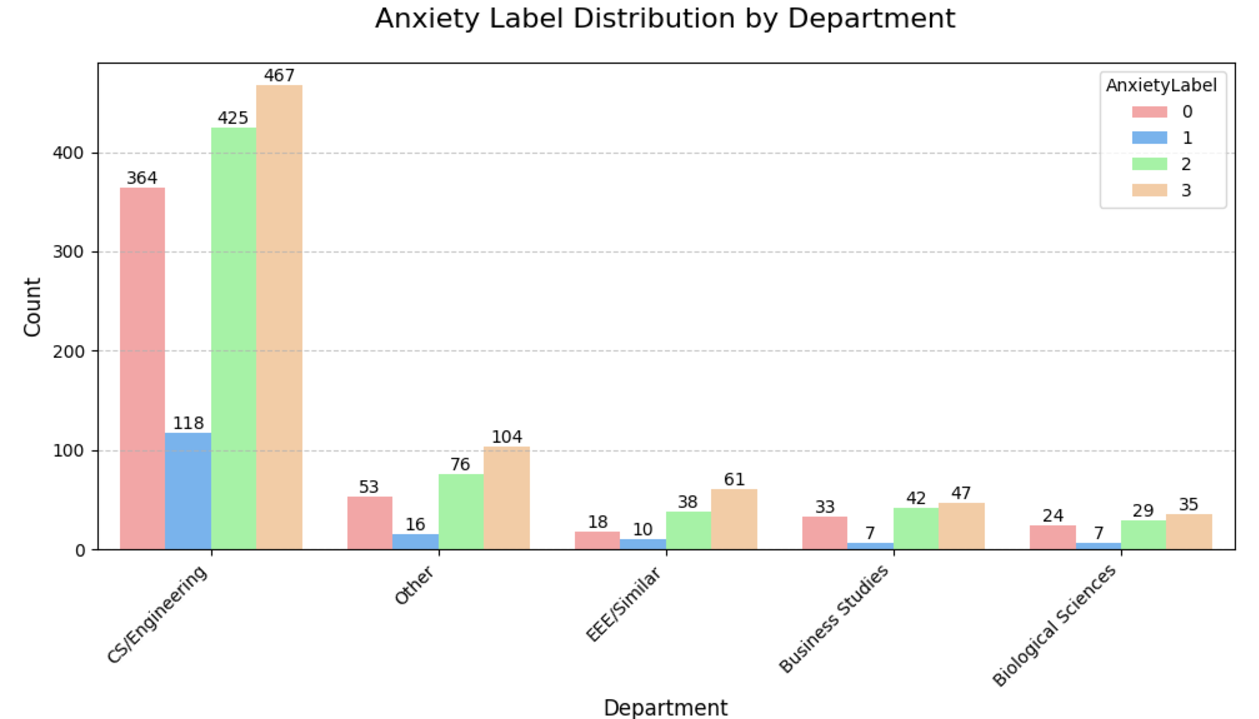


Fig.8. Anxiety label Vs Department

The above Fig.8. shows Anxiety label Vs Department. Here we can anxiety that CSE Engineering is suffering more from Anxiety.

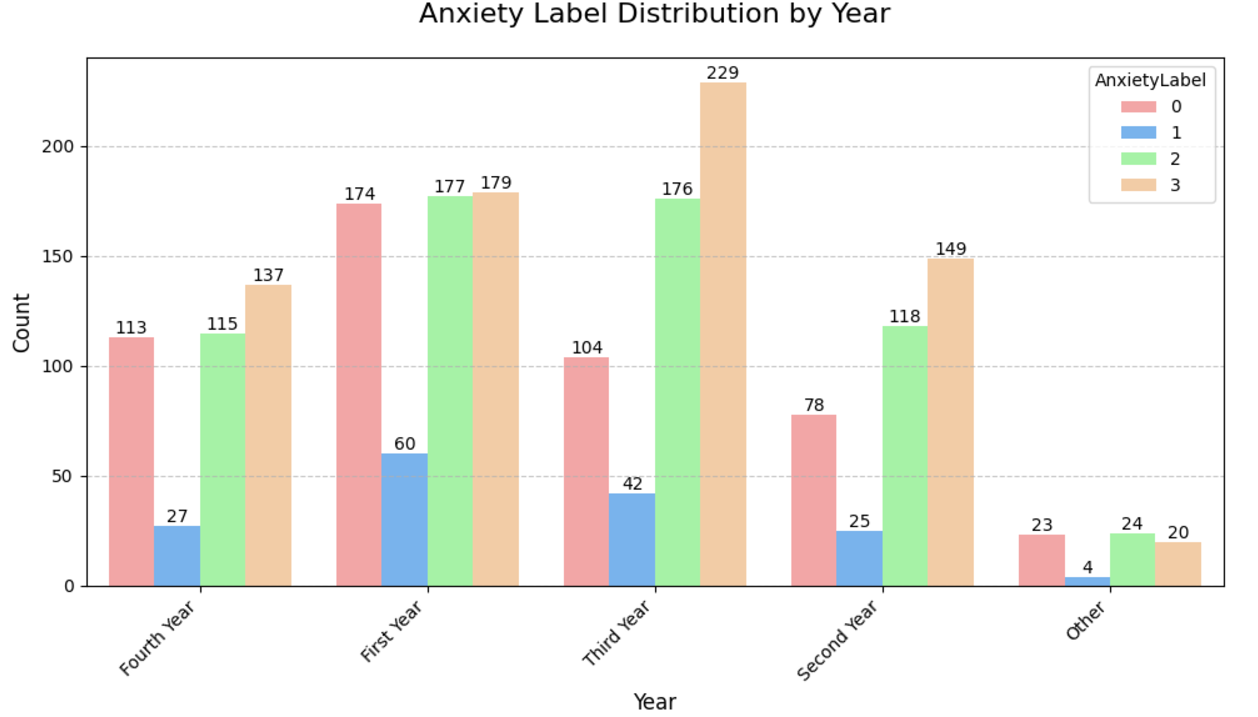


Fig.9. Anxiety label Vs Year

The above Fig.9. shows Anxiety label Vs Year. Here we can identify that Third-Year students is suffering more from Anxiety.

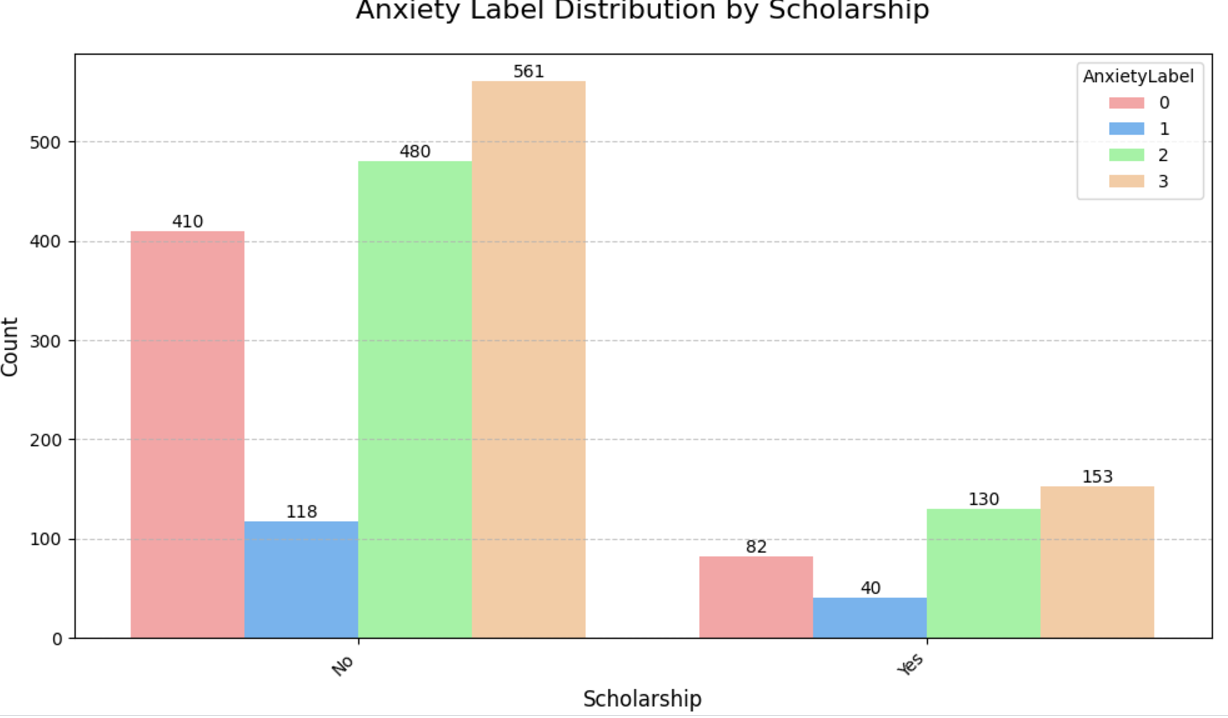


Fig.10. Anxiety label Vs scholarship

The above Fig.10. shows Anxiety label Vs scholarship. Here we can identify that students who have not received Scholarship has suffered more from Anxiety.

# METHODOLOGY

In this project, we employed three machine learning models--- Logistic Regression, K Nearest Neighbors and Nave Bayes--- to analyze and predict the outcome of our data set. Each model was implemented and trained using standard techniques and parameters optimized through experimentation.

## Logistic Regression

Logistic regression is a powerful tool for predicting anxiety levels in students based on various data sets containing features such as Nervous, Worrying, Relaxing, Irritated, Overthinking, Restless, Afraid and Anxiety label, Upset, Control, Stressed, Cope, Confident, Down, Sleep, Energy, Appetite, Self-worth, Concentration, Movement, Depression Level patterns and survey response by modeling the probability of students experiencing anxiety, logistic regression uses these features to establish relationships with the target variable. Its probabilistic framework allows it to estimate the likelihood of anxiety, enabling precise classification based on chosen thresholds. Logistic regression is interpretable, meaning the influence of individual factor on anxiety can be quantified through feature coefficient. This makes it particularly valuable in education and psychological studies, as it not only predicts outcome, but also provides actionable insight for intervention strategies. By effectively handling binary classification tasks, Logistic Regression is a robust choice of anxiety prediction across diverse student datasets.

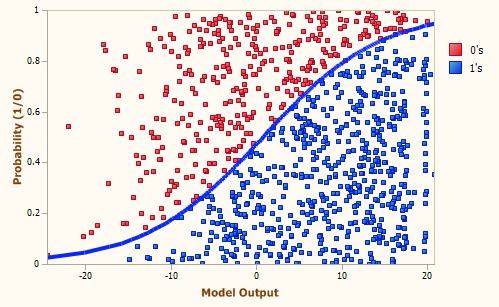
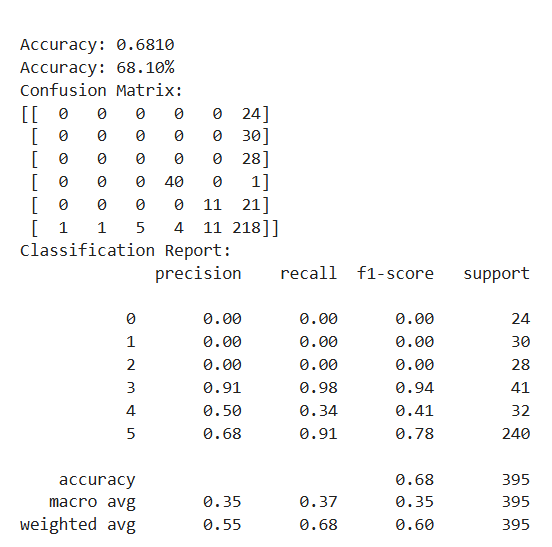


Fig.10. Logistic Regression

After applying Logistic Regression on the trained dataset, we have calculated different parameter i.e. Precision, Recall, F1-Score, Support and Accuracy of anxiety as shown in Fig 11. We have got an Accuracy of 68.10% for Logistic Regression.



## K Nearest Neighbors (K-NN)

The K-Nearest Neighbors (KNN) algorithm is a straightforward and adaptable machine learning technique that can be used to predict student anxiety based on behavioral, academic, and personal information. This algorithm evaluates the similarity between a student’s data—such as academic scores, sleep patterns, survey results, and social activities—and that of other students in the dataset. It determines the kk nearest neighbors by calculating distances (e.g., using Euclidean distance) to find students with similar characteristics. Anxiety is then predicted based on the predominant class (e.g., "anxious" or "not anxious") among these nearest neighbors. KNN is particularly effective for datasets with complex, non-linear relationships, as it does not rely on any assumptions about data distribution. By selecting an appropriate kk value, the model can balance noise reduction and accuracy, making it a valuable tool for identifying anxiety in diverse student datasets.

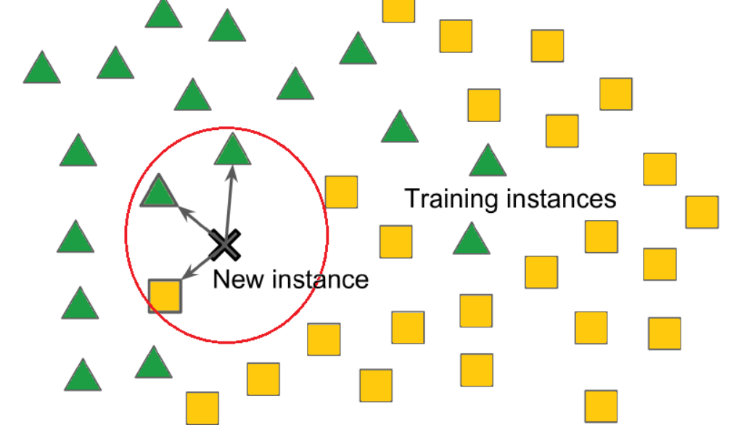
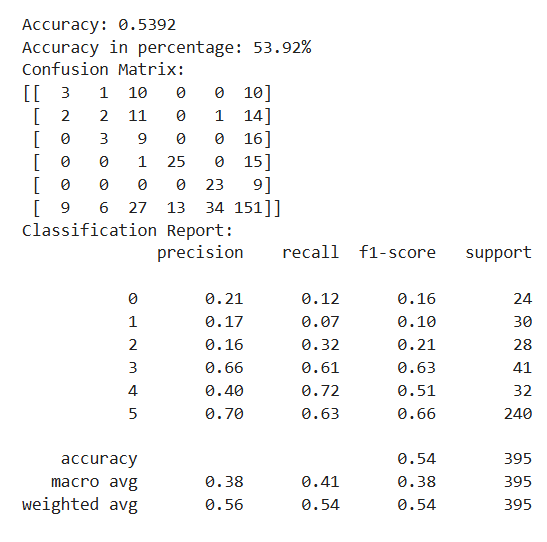


Fig.11. K-Nearest Neighbors (KNN)

After applying K-Nearest Neighbors (KNN) on the trained dataset, we have calculated different parameter i.e. Precision, Recall, F1-Score, Support and Accuracy of anxiety as shown in Fig 11. We have got an Accuracy of 53.92% for K-Nearest Neighbors (KNN).



## Naïve Bayes

Naive Bayes is a probabilistic machine learning algorithm that can be effectively used to predict anxiety in students by analyzing various features, such as academic performance, lifestyle habits, and survey responses. The model applies Bayes’ theorem, assuming that all features are conditionally independent, to calculate the probability of a student being "anxious" or "not anxious" based on the input data. Despite its simplicity, Naive Bayes performs well for classification tasks, especially with categorical data. By learning the likelihood of anxiety based on individual features and combining them to make predictions, it provides a fast and efficient way to analyze student data. This algorithm is particularly useful for datasets with clear class distinctions and works well even with smaller datasets, making it an excellent choice for predicting anxiety among students in a structured and interpretable manner.

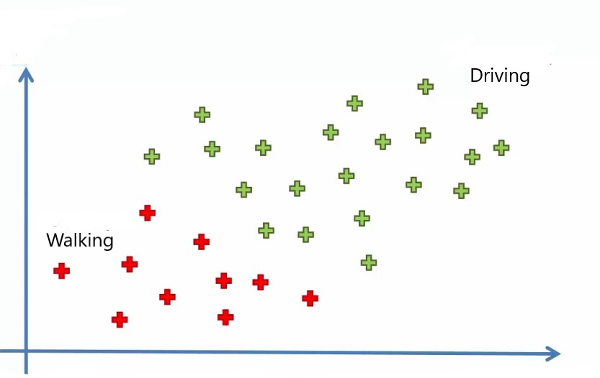
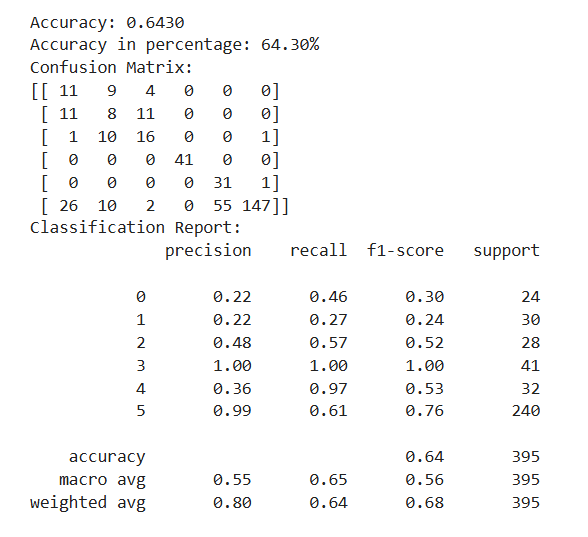
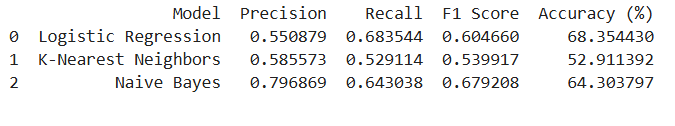


Fig.12.Naive Bayes

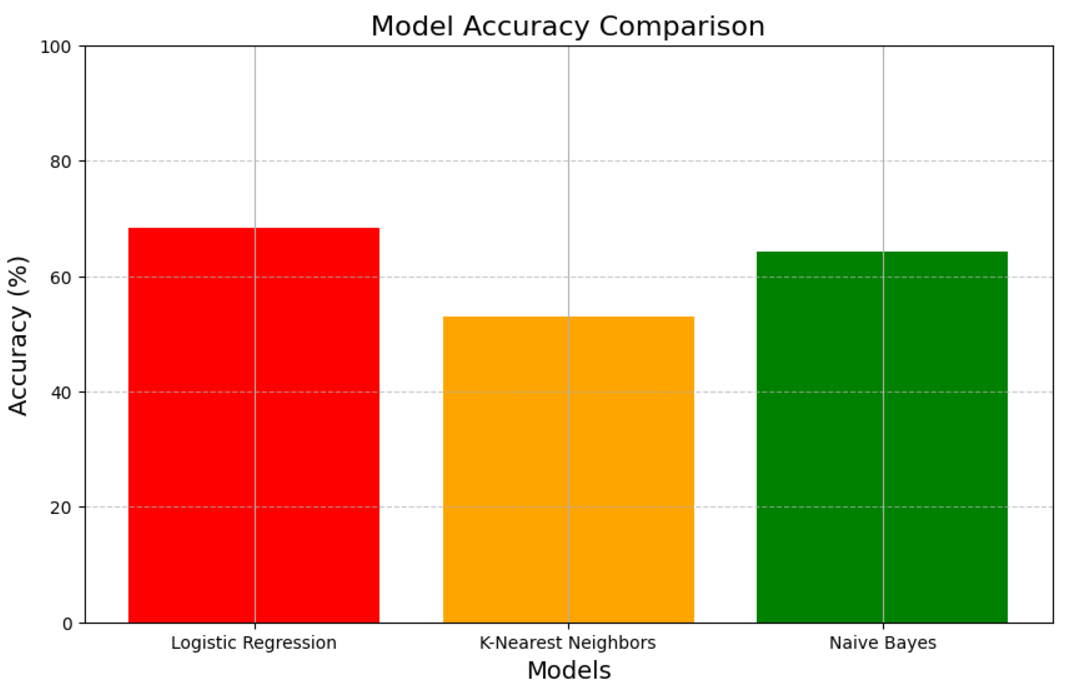
After applying Naive Bayes on the trained dataset, we have calculated different parameter i.e. Precision, Recall, F1-Score, Support and Accuracy of anxiety as shown in Fig 11. We have got an Accuracy of 64.30% for Naive Bayes.



After identifying the accuracy of each model used to train the dataset, we found that Logistic Regression has given an Accuracy of 68.10% which is highest among all the models we have trained. So after training the dataset we have identified Logistic Regression is best suited for our Project



We have showcased the Accuracy of anxiety of students of different dataset in the form of table and bar graph as shown in Fig.11.



# conclusion

Through this project on anxiety prediction in students’ life, we have gained significant insight into the practical application of machine learning techniques. The project began with the data processing, where we learned to identify and handle duplicate values, manage unique values and address outliers by analyzing and where necessary, removing them. Data cleaning and preparation were crucial steps, ensuring the data set was ready for analysis. Additionally, we explored the use of a correlation matrix to understand relationships between features and refine the data set for training.

Following preprocessing, we implemented an evaluated three machine learning models, that is, Logistic Regression, K Nearest Neighbors and Naïve Bayes. Each model brought unique strength, but the evaluation revealed that Logistic Regression outperformed the others in terms of predicting accuracy of anxiety levels. The result underscores the

importance of model selection and the value of experimentation in achieving reliable outcomes.

# reference

1. ***Bharadwaj, S., Pai, R. M., & Balasubramanian, V. (2021)****. "A Machine Learning Approach for Predicting Anxiety Levels in College Students Based on Social and Academic Factors." International Journal of Medical Informatics, 153, 104514.*
2. **Yang, Y., & Ju, Y. (2019)**. *"Predicting Adolescent Anxiety and Depression Based on Interaction Behavior with Social Media: A Machine Learning Approach."* Computers in Human Behavior, 98, 1-10.
3. **Lin, C. J., Hsieh, H. F., & Wang, Y. J. (2020)**. *"Using Machine Learning to Predict Student Stress and Anxiety in Daily Life."* Sensors, 20(20), 5787.
4. **Acharya, U. R., Sudarshan, B., & Rajendra, V. (2021)**. *"Utilizing Machine Learning for Mental Health Diagnosis Based on Daily Activity Patterns."* IEEE Transactions on Neural Systems and Rehabilitation Engineering, 29, 1043-1052.
5. Kroenke, K., Spitzer, R. L., Williams, J. B. (2001) "The PHQ‐9: validity of a brief depression severity measure. "Journal of general internalmedicine 16 (9): 606-613.

|  |
| --- |
|  |
|  |
|  |  |