NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY School of Electrical Engineering and Computer Sciences

Research Report

"Comparative Analysis of Machine Learning Models for Predicting Stress Levels among Undergraduate Students at NUST, Pakistan"

Submitted to: Ms. Ayesha Habib

Submitted by:

Name:	CMS ID:
Muhammad Furqan Raza	460535
Tamkeen Sara	474585
Rimsha Mahmood	455080
Attiqa Bano	473781
Hajra Asif	454056

Date of Submission: May 16th, 2025

Class: BS-DS-1A

Abstract

This research examines how machine learning models can be used to predict undergraduate students' stress levels at Pakistan's National University of Sciences and Technology (NUST). To determine the *academic, personal, and demographic factors* impacting stress, we gathered survey data from *212 undergraduate students* in a variety of fields using a quantitative cross-sectional approach. The research addressed three objectives: *to compare model performance, identify significant stressors, and train machine learning models to categorize stress levels.* The majority of students have moderate to high levels of stress, according to the results (45 low, 95 medium, 71 high). The best performance of all the ML algorithms examined was shown by *Logistic Regression with SMOTE balancing technique (97.30% accuracy)*, which was followed by *Random Forest and SVM (89.19% each)*. The research revealed *anxiety level* (r=+0.745) and *social isolation* (r=+0.454) as the most significant predictors of elevated stress, although academic workload, peer pressure, coffee use, and social media usage had minor positive correlations. On the other hand, stress was negatively correlated with extroverted personality qualities, academic satisfaction, regular exercise, enough sleep, and a nutritious diet. The findings underscore the urgent need for universities to implement targeted mental health interventions. At the individual level, we suggest peer counseling circles, stress awareness courses, individualized mental health tracking, and digital detox periods. Wearable physiological data, a more diverse population, multi-label categorization using deep learning methods, and real-time monitoring systems for early stress detection are all important areas for future research. This study demonstrates the potential of ML algorithms as valuable tools for the early identification of stress among Pakistani university students.

Acknowledgements

We would like to express our heartfelt gratitude to all those who supported us throughout the course of this research. Special thanks go to our peers and colleagues for their cooperation and moral support. Above all, we are profoundly grateful to our families, whose unwavering faith, patience, and encouragement have been a constant source of strength and inspiration.

Table of Contents

Chapter 1: Introduction

This chapter provides the foundational concept of stress and its association with university students. It then researches how machine learning algorithms can be used to identify stress effectively. It then defines the core objectives of this research, underlines its contributions to the academic and social community, and concludes with the scope and limitations of this study.

University life is delightful, but is often accompanied by workload and emotional instability. That is why it is very important to use data-driven Machine Learning methods to understand and help the students manage their stress strategically.

1.1 Background of the Study

Stress is a universal *mental health condition* that occurs among a large number of individuals globally. It involves recurrent feelings of sadness, loss of interest, and impaired function, resulting in a significant decrease in overall well-being and quality of life.

Stress is a global phenomenon that profoundly affects individuals' psychological and physical health. Among all affected populations, university students are especially vulnerable because of academic stress, social transitions, and career uncertainties. While students cope with hectic timetables, competitive settings, and life-changing choices, their *susceptibility to stress* has become an alarming issue, especially in countries like *Pakistan*, where mental health support is still developing.

Recent developments in Artificial Intelligence (AI) and Machine Learning (ML) offer potential tools for the detection of mental health risks like anxiety, depression, and chronic stress. These models can analyze complex behavioral, physiological, and demographic data to provide predictive insights. Research has shown that models like *decision trees, support vector machines (SVM)*, and *ensemble techniques* can accurately categorize mental distress levels.

Studies also indicate the growing success of using machine learning for numerous domains, including the *healthcare domain.* In this research, we use ML techniques for survey-based mental health predictions.

By learning from diverse attributes, which range from screen time to sleep patterns, these models can help flag students with high stress levels who might benefit from early mental health support or intervention.

In this research, we aim to compare the *performance of multiple ML models* to predict stress levels among undergraduate students from *public and private universities* in Islamabad, Pakistan. The goal of this research is to find the *best ML model* for predicting stress levels among students in Pakistan and to extract the key attributes impacting the stress levels. By conducting the study within Islamabad, we assure the research is rooted in the local educational environment and reflects the cultural and psychological conditions of Pakistani students.

1.2 Objectives of the Study

- To *train* Machine Learning Models on data collected from university students of Islamabad and classify them as Low, Medium, or High Stress Category

- To *compare* the performance of different machine learning models (Linear Regression, Decision Tree, Support Vector Machine(SVM), Random Forest) in predicting Students' Stress Levels.

- To *identify* the most influential demographic, academic, and personal factors associated with stress among students.

1.3 Research Questions

This study is guided by the following research questions:

- How accurately can machine learning models predict student stress levels based on survey-based data?

- Which ML algorithm performs best in terms of accuracy, precision, recall, and F1 score in this context?

- What *demographic, academic,* and *personal* factors have a prominent contribution to stress among undergraduate students in Islamabad?

1.4 Significance of the Study

This research is significant in multiple areas:

It contributes to the *expanding research on mental health prediction* using AI/ML, specifically in the educational context of Pakistan, as there is a significant research gap regarding this topic.

By understanding key features affecting the stress and accurate predictors, universities can design more *targeted mental health interventions* to create a balance among academic pressure, student counseling access, and extracurricular activities.

The insights gained will not only benefit NUST but will serve as a model for mental health screening initiatives across other universities in Pakistan and similar educational ecosystems to shape targeted recommendations and support systems for university students. It will help reduce their stress level by *implementing appropriate strategies* and ensuring students' mental well-being in academic institutions.

It will also highlight the importance, efficiency, and *impact of Machine Learning techniques* on real-world problems by showcasing their practical utility in the domain of mental health through the "Stress Prediction," which will encourage and increase confidence in using *data-driven ML solutions* for diverse societal challenges in Pakistan.

1.5 Scope and Delimitation

Financial and time constraints limit the creation of dedicated datasets for stress prediction. Due to the infeasibility of data collection from all universities of Islamabad in a limited time, this study will focus on undergraduate students enrolled in four universities in Islamabad—two private and two public. National University of Science and Technology (*NUST*) and National University of Computer and Emerging Sciences (*NUCES*) will be private institutions included. Public Institutions enrolled in this study are *Quaid-e-Azam University* and *ARID Agriculture University*. The reason these universities have been chosen roots from having easier access to them and their ability to forecast different factors affecting stress levels among students in both public and private sectors.

Students will be categorized into three ordinal levels of Stress: *Low, Medium*, and *High*, based on students' responses to a structured questionnaire. The predictive attributes will be grouped into three categories:

- Demographic Factors (e.g., gender, BMI, residential status, diet, personality type)
- Academic Indicators (e.g., exam period, GPA, satisfaction with teachers)
- Personal Lifestyle (e.g., screen time, sleep hours, mental health history)

Delimitations include the focus on *undergraduate* students only and reliance on self-reported survey responses, which may introduce a degree of subjectivity. Moreover, as a future improvement, this study has the potential to generalize its findings on university students all over Pakistan with an extensive dataset containing more relevant features and improved ML models.

By merging the strengths of data science and behavioral psychology, this study hopes to not only compare predictive models' performance but also highlight meaningful insights that can shape student support systems in higher education institutions.

This chapter had information regarding the background of stress and its effects, the objectives, and the significance of our research. By merging the strengths of data science and behavioral psychology, this study hopes to not only compare predictive ML models but also highlight the key features affecting stress in university students and find the meaningful insights that can shape student support systems in higher education institutions. It will help diagnose stress and depression at an early stage and help the undergraduates cope with it in a better way.

Chapter 2: Literature Review

In recent years, the intersection of machine learning (ML) and mental health prediction has gained considerable traction, particularly in addressing student stress. Researchers have developed various predictive models and analytical frameworks to identify stress levels among university students using survey-based and behavioral data.

Bokolo and Liu (2023) created a machine learning framework to predict stress levels among university students by including psychological, academic, and lifestyle variables. Their study emphasized the effectiveness of prediction algorithms like Logistic Regression and Random Forest in classifying stress. Likewise, *Alghamdi et al. (2022)* presented an extensive review of machine learning techniques used in stress detection, highlighting greater use of supervised learning models like Support Vector Machines, Decision Trees, and Naïve Bayes in various contexts.

Yadav and Singh (2020) analysed student stress using ML algorithms and concluded that sleep patterns, academic workload, and peer pressure significantly influenced stress predictions. *Rashid et al. (2021)* also confirmed the utility of ML by using classification algorithms to predict psychological stress levels and highlighted the importance of preprocessing and feature engineering.

Vaibhav and Vyas (2021) introduced a hybrid model combining ensemble methods and neural networks, significantly improving mental health prediction accuracy. In line with this, *Alahmadi and Hussain (2022)* developed a predictive model focusing specifically on mental health risk assessment in university students, demonstrating the high sensitivity of ML techniques in early detection.

Salehinejad and Beheshti (2019) explored deep learning applications in mental health prediction and revealed that neural networks could capture complex behavioral patterns more effectively than traditional models. *Shatte et al. (2019)* performed a scoping review to map out machine learning's role in mental health, underscoring that while many models exist, their real-world application in educational settings is still under development.

Choudhury and Basu (2020) utilized behavioral traits and supervised learning for predicting student stress, highlighting the correlation between digital behavior and mental health. *Nguyen and Nguyen (2021)* employed survey-based datasets and confirmed that even simple ML models can yield substantial accuracy when properly tuned with relevant features.

Arora and Malik (2022) conducted a comparative analysis of classifiers for stress level prediction and found Random Forest to outperform others in accuracy and robustness. *Kaur and Sharma (2020)* offered a comprehensive summary of machine learning use in predicting mental health outcomes and stressed the need for further contextual research, particularly in developing nations.

Finally, *Muhammad and Habib (2023)* have suggested an AI-based stress detection system for Pakistani students with a focus on region-based models taking into account cultural and education-specific stressors in the region.

Collectively, these studies provide the groundwork for the development of successful ML models that are specific to the academic and psychological contexts of university students, especially in countries such as Pakistan, where mental health infrastructure is being established.

Chapter 3: Methodology

This section outlines the methodology employed in this study to detect stress levels among students using machine learning (ML) techniques. The approach encompasses research design, data collection, preprocessing, feature engineering, and the selection and training of ML algorithms.

3.1 Research Design

This research will employ a quantitative, survey-based descriptive cross-sectional design that will involve the non-experimental collection of data from undergraduate students aged between 18 to 25 across four different universities in Islamabad. This design will enable efficient data collection from a diverse group of participants over a limited time frame, allowing for the analysis of stress levels and their association with various academic factors.

3.2 Population and Sample Size

This study will focus on students from the National University of Sciences and Technology (NUST). We will aim to collect responses from 200 to 250 students, for which participants will be approached via university mailing lists. It will be made to include students from different academic years, departments, and backgrounds. The sample will be chosen using random sampling. To ensure the results are reliable, we will calculate the required number of participants using a 95% confidence level, a 5% margin of error, and a 50% response rate assumption to cover a wide range of possible outcomes.

3.3 Data Collection

Data will be collected using a structured digital questionnaire that will be disseminated through Google Forms. The survey will be designed to gather information across three different domains.

Section 1 (Demographic Information)	Section 2 (Academic-related Variables)	Section 3 (Personal and Psychological Variables)
Age, Gender, Diet Quality, Personality Type, Discipline, Residential Status (Day-scholar/Hostelite), Financial Status	Academic Workload, Study Hours, CGPA, Peer Pressure, Teacher Satisfaction Level, Future Career Concerns	Sleep Patterns, Physical Activity, Social Support, Screen Time, Coping Mechanisms, Mental Health History

Table 1.1

These factors will be selected based on prior literature highlighting their potential impact on stress and mental health among university students.

3.4 Machine Learning Models

In this section, two machine learning models will be selected for the comparative analysis: Logistic Regression (LR) and Random Forests (RF) will be trained and optimized using the same input features identified through subgroup analysis of university students with and without severe mental distress.

Logistic regression (LR) is a widely used linear classifier. It models the relationship between the independent variables and the probability of a binary outcome, making it suitable for classification tasks. By employing multinomial approaches, LR can be adapted to handle multi-class problems, such as our categorization of stress levels. In our study, we will develop and train an LR model on the preprocessed data to classify stress levels among students into three categories: low, medium, and high *(Bokolo & Liu, 2023)*. Figure 1.1(a) demonstrates the working of LR for classifying 3 classes. *(Analytics Vidhya, 2025, May 16)*

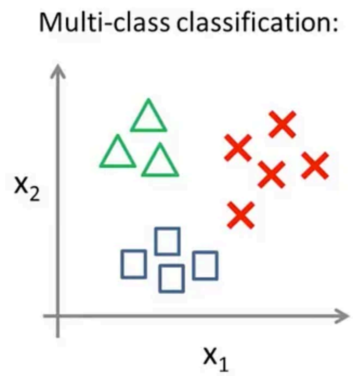


Fig. 1.1(a)

Random forests (RF) are ensemble learning algorithms that combine multiple decision trees to make predictions. They are capable of capturing complex interactions and non-linear relationships in the data. We will employ random forests to leverage the ensemble of decision trees for classifying stress levels based on various survey-based features related to student stress (Bokolo & Liu, 2023). The following Figure 1.1(b) explains the working of RF for 3-class classification. *(A schematic depiction of the classification flow of Random Forest with three decisions. ResearchGate, 2023. Fig. 3)*

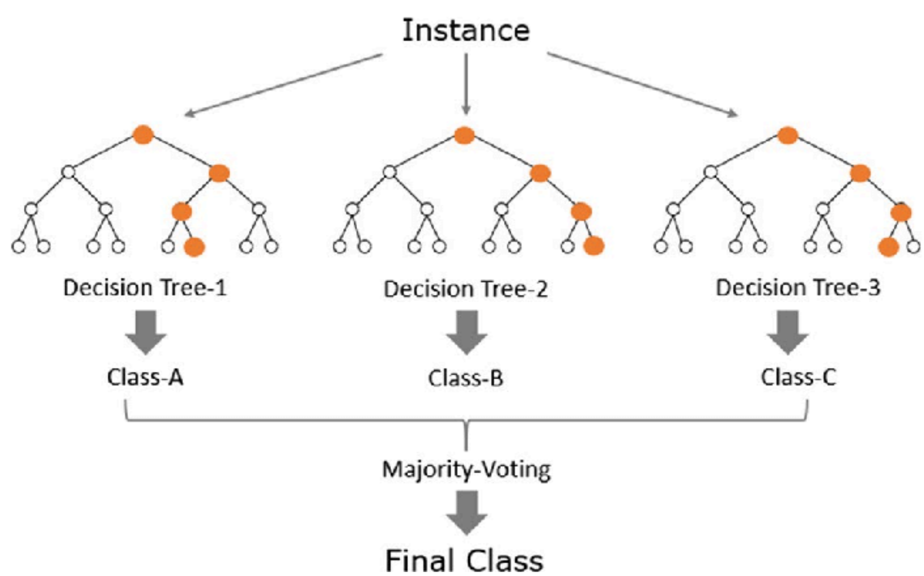


Fig. 1.1(b)

Support Vector Machine (SVM) is a supervised learning technique for classification and regression. However, in machine learning, SVMs' primary application is in the classification task. The principle behind the SVM algorithm is to find the optimal decision boundary/hyperplane that divides the n-dimensional space into classes, making it easier for a new dataset to categorize into a new class. Fig. 1.1(c) represents the working SVM model for a three-class classification *(Emine Cengil, 2021)*. Multiple decision boundaries can differentiate the three data groups, and the primary purpose of SVM is to choose the optimal one among all the decision boundaries. This can be done by selecting the largest possible margin or separation between the two classes of data. *(Pradhan, B. K., Saha, S., Pal, K., & Banerjee, I. , 2023).*

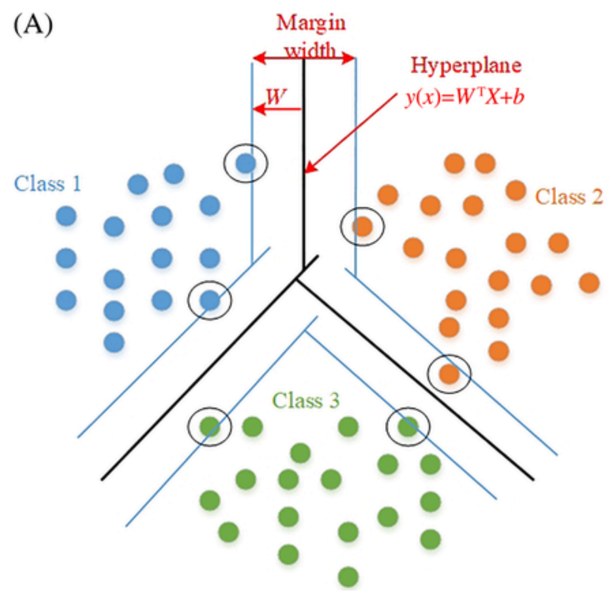


Fig. 1.1(c)

eXtreme Gradient Boosting (XGBoost) is a decision tree ensemble based on gradient boosting designed to be highly scalable. It builds an additive expansion of the objective function by minimizing a loss function *(Gonzalo Martınez-Munoz, 2019)*. It uses decision trees as its base learners, combining them sequentially to improve the model's performance. Each new tree is trained to correct the errors made by the previous tree, and this process is called boosting *(GeeksforGeeks)*. Figure 1.1(d) demonstrates the working of XGBoost for classifications *(Fraidoon Omarzai, 2024)*

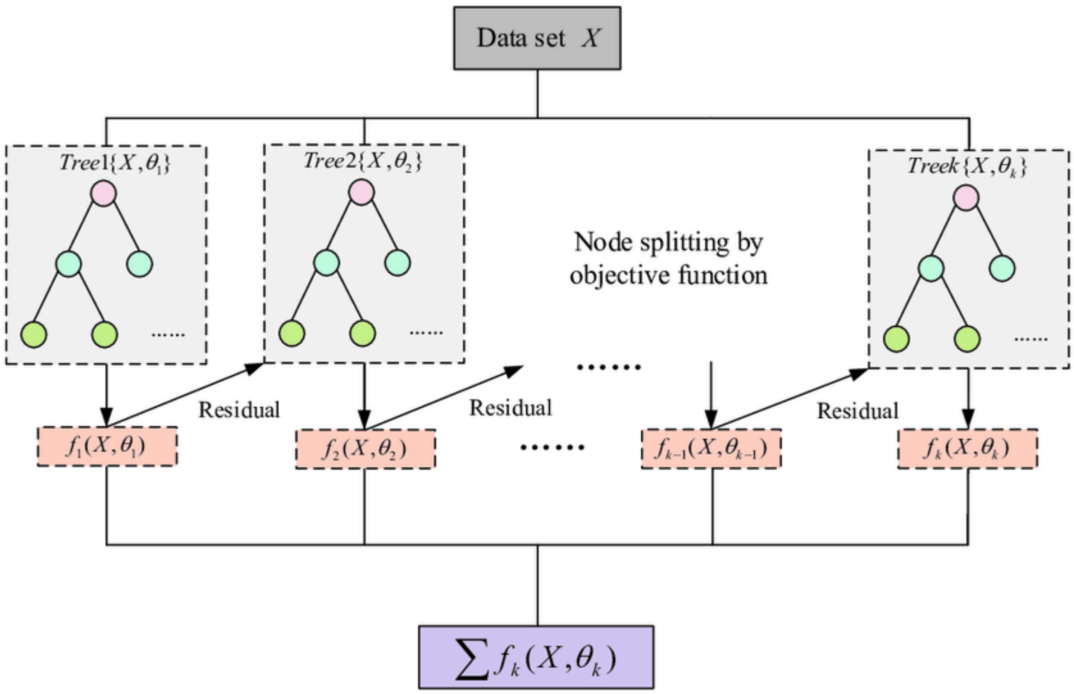


Fig. 1.1(d)

K-Nearest Neighbor (KNN) is a supervised learning tool used in regression and classification problems. In the training phase, KNN uses a multi-dimensional feature vector space that assigns a class label to each training sample. KNN generates classification results by storing all the available cases and stratifying new classes based on a similarity measure (distance functions). *(Sulaiman Khan, 2018)*. The following Figure 1.1(e) explains the working of KNN for 3-class classification. *(Sachinsoni, 2023)*

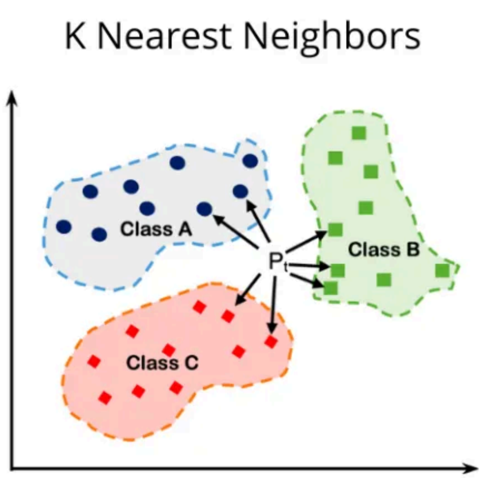


Fig. 1.1(e)

Proposed Framework

The custom-designed ML pipeline suggested in this work is shown in Figure 1. The authors constructed the design and flowchart to best fit the dataset and goals of the project.

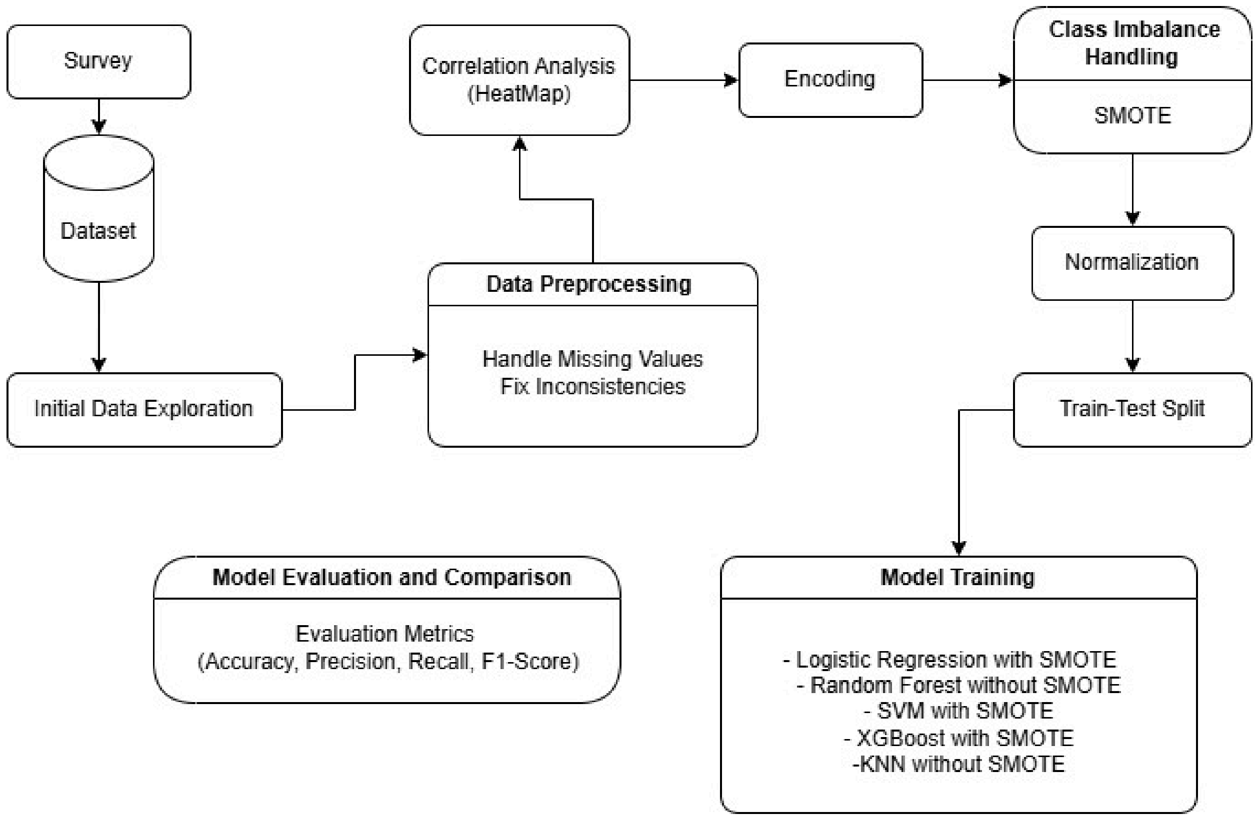


Fig. 1.2

Synthetic Minority Over-sampling Technique (SMOTE) is a powerful method used to handle class imbalance in datasets. SMOTE handles this issue by generating samples of minority classes to make the class distribution balanced. SMOTE works by generating synthetic examples in the feature space of the minority class. *(GeeksforGeeks)*. Here's the breakdown of SMOTE:

1. **Identify Minority Class:** SMOTE starts by detecting the underrepresented class in the dataset.

2. **Find Nearest Neighbors:** It then finds *k* nearest neighbors for each minority instance.

3. **Generate Synthetic Samples:** New data points are created between the instance and a randomly chosen neighbor.

4. **Control Oversampling:** The number of synthetic samples is determined by a user-defined oversampling ratio.

5. **Balance the Dataset:** This process repeats for all minority instances, resulting in a more balanced dataset.

3.5 Validity and Reliability

We will apply both LR and RF using *k-fold cross-validation* for accurate results. This approach splits the dataset and keeps the stress level categories balanced in each fold, which helps prevent overfitting. The models will be evaluated using standard metrics, i.e., accuracy, precision, Recall, and F1 score. The thorough validation process will help ensure that the findings are reliable and can be applied to different student groups without any bias.

3.6 Limitations of the Study

The sample may not accurately represent the broader student population, affecting the applicability of the findings. The study may have overlooked important stress-related factors due to the short data collection period. Additionally, financial limitations and challenges in recruiting participants may have restricted the study's scope and diversity. Furthermore, results may contain errors due to self-reported data and potential researcher bias. Future studies should aim for larger sample sizes, extended study durations, and increased funding to enhance generalizability and reliability.

3.7 Ethical Considerations

Notably, data collection will be anonymous. All the participants will be informed about the study's purpose, and their responses will be gathered only after consent.

3.8 Summary

The study will use a quantitative research design with closed-ended surveys to collect data. The sample size will be 200-250, selected using random sampling. Data will be analyzed using ML techniques and relevant statistical methods in Python. All procedures will follow strict ethical guidelines, including informed consent and data confidentiality. The methodology will be well-designed to ensure the validity, reliability, and quality of the collected data.

Chapter 4: Results and Findings

This chapter presents the results of the survey administered to 250 university students. Each section provides detailed insights and interpretations, along with clearly defined placeholders for graphical visualizations.

4.1 Overview of the Data

The analysis includes one main respondent group:

- NUST UG Students: 212 respondents, mainly undergraduate students from various disciplines, including Engineering and Technology, IT, Business, Medical Sciences, Humanities, and Natural Sciences.

This sample, gathered via a quantitative survey, provides an in-depth analysis of stress levels among NUST students, facilitating the development and comparison of ML models for stress prediction.

The dataset is inherently imbalanced, with each class of stress (low, medium, and high) exhibiting different distributions. This issue was addressed using the SMOTE technique, as outlined in Chapter 3.

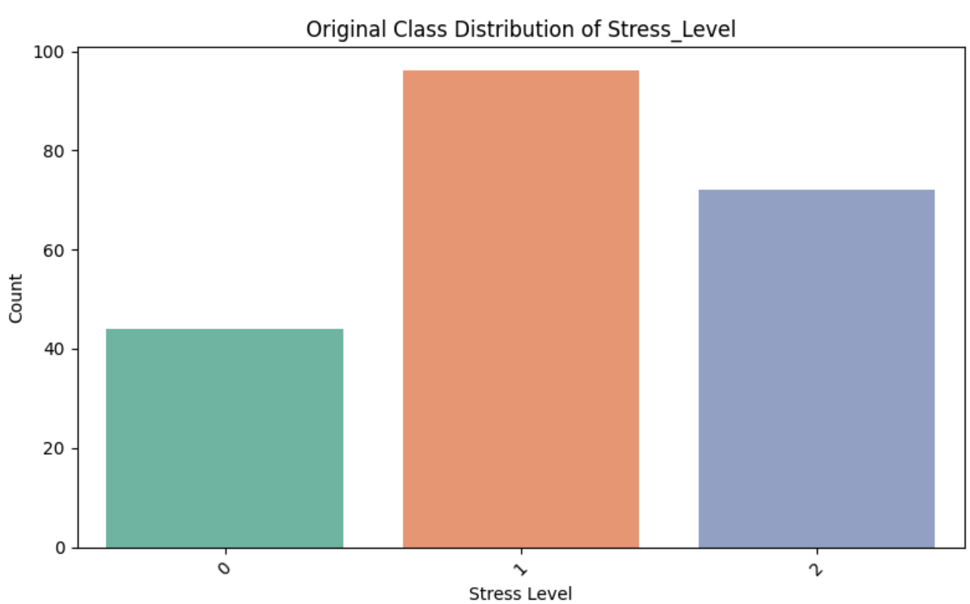


Fig. 2 (a)

4.2 Descriptive Analysis of Survey Responses

4.2.1 Summary Statistics

Variable	Count	Mean	Std Dev	Min	Q1 (25%)	Median (50%)	Q3 (75%)	Max
Age	212	21.17	2.05	18	20	21	22	29
Personality_Type	212	0.38	0.31	0	0	0.5	0.5	1
Is_Hostellite	212	0.6	0.49	0	0	1	1	1
Diet_Quality	212	2.49	0.96	1	2	2	3	5
Financial_Condition	212	3.59	1	1	3	4	4	5
Family_Support	212	4.43	0.91	1	4	5	5	5
Side_Hustle	212	0.25	0.43	0	0	0	0	1
Study_Load	212	3.96	0.91	1	3	4	5	5
Satisfaction_Level_with_Instructors	212	2.57	1.04	1	2	3	3	5
Career_Concern	212	4.36	0.75	2	4	5	5	5
Socially_Isolated	212	2.96	1.23	1	2	3	4	5
Social_Media_Usage	212	3.77	1.01	1	3	4	5	5
Reels_Engagement	212	3.86	1.15	1	3	4	5	5
Peer_Communication	212	3	1.25	1	2	3	4	5
Desire_to_travel_home/hostel	212	4.11	1.13	1	4	4	5	5
Caffine_Intake	212	2.71	1.47	1	1	2	4	5
Substance_Intake	212	1.38	1.04	1	1	1	1	5
Societies_Participation	212	2.18	1.37	1	1	2	3	5
Social_Events_Participation	212	2.44	1.3	1	1	2	3	5
Stress_Level	212	1.13	0.73	0	1	1	2	2

Fig. 2 (b)

4.2.2 Demographic Features

1. **Age**

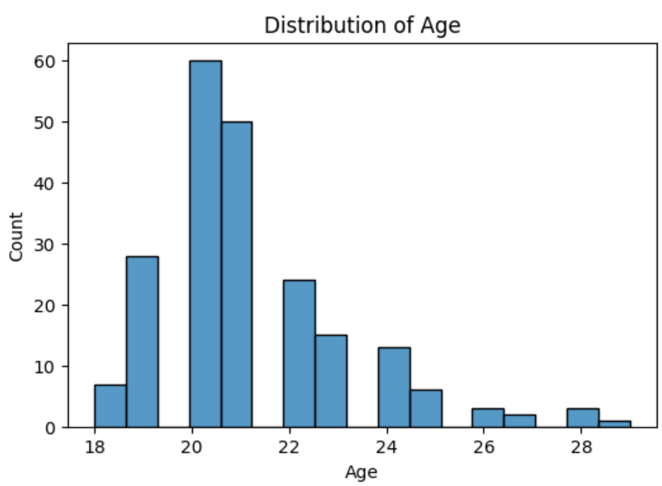


Fig. 3.1 (a)

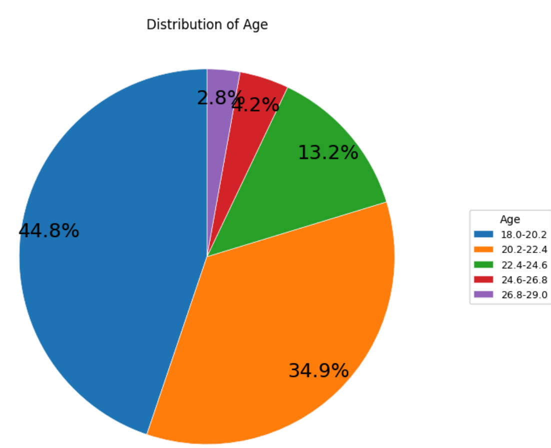


Fig. 3.1(b)

The distribution shows 78% of students are aged 18–22, complementing the mean (21.17) and median (21). Only 5% are above 25, indicating a young population.

2. Personality Type

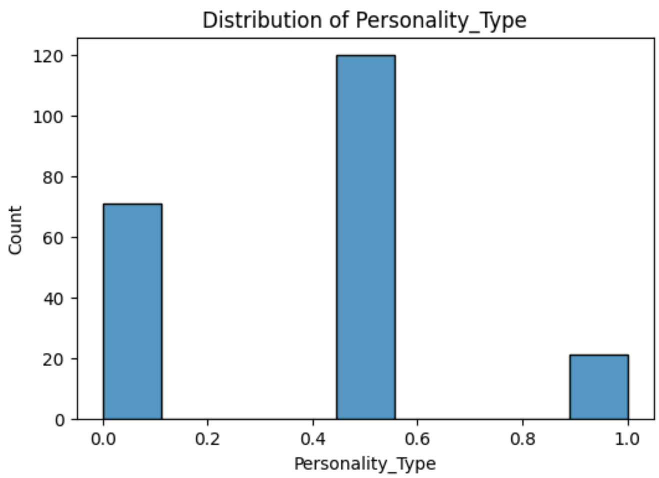


Fig. 3.1(c)

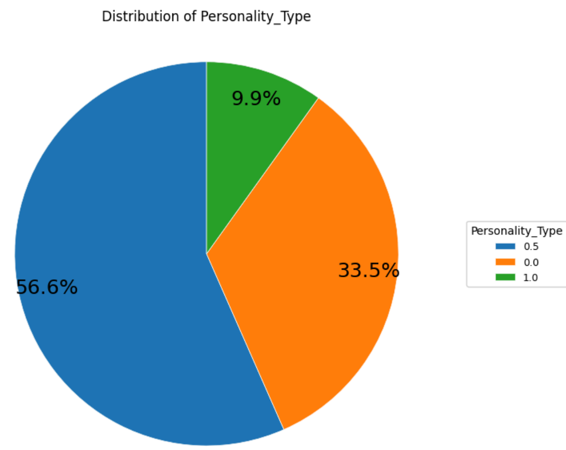


Fig. 3.1(d)

The distribution shows that about 34% are introverts (0), 56% are ambiverts (0.5), and only 10% are extroverts (1).

3. Gender

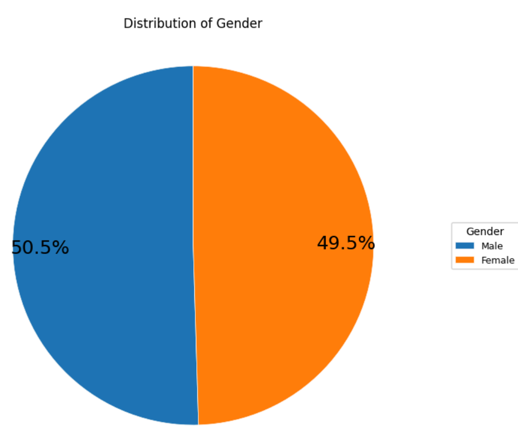


Fig. 3.1(e)

The distribution reveals that the number of male and female respondents is almost balanced, aligning with the initial goal.

4. Disciplines

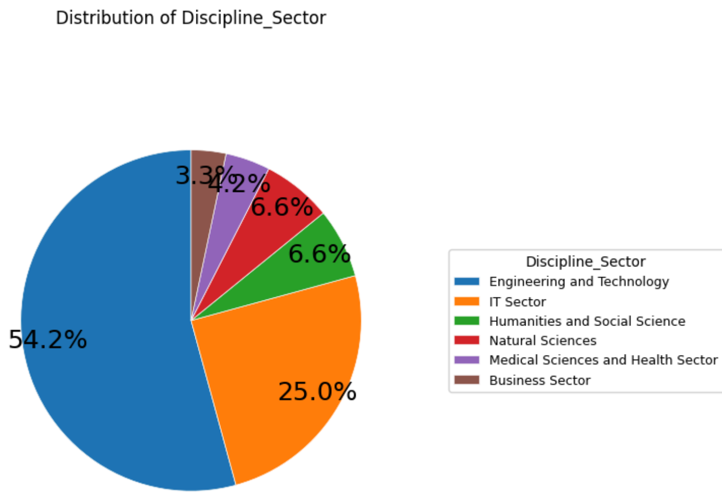


Fig. 3.1(f)

Students drawn from Engineering, IT, Business, and Social Sciences—reflecting the random sampling strategy (see Chap. 3).

5. Residential Status

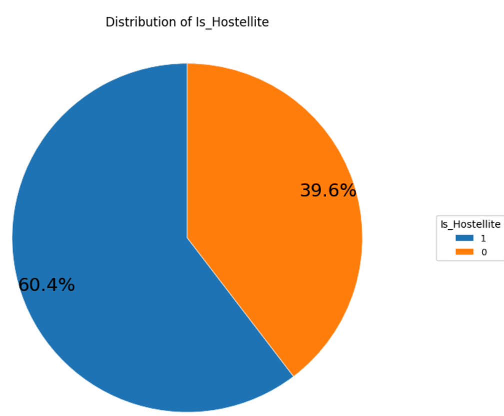


Fig. 3.1(g)

About 60% of respondents are hostelites (1), and 40% are day scholars (0).

6. Diet Quality

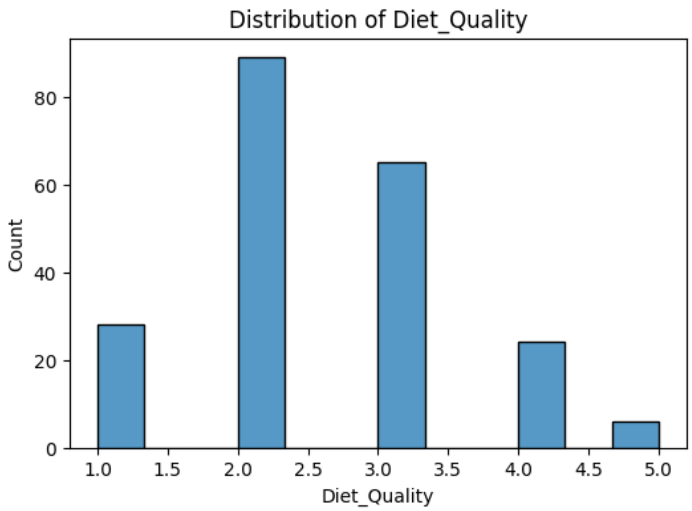


Fig. 3.1(h)

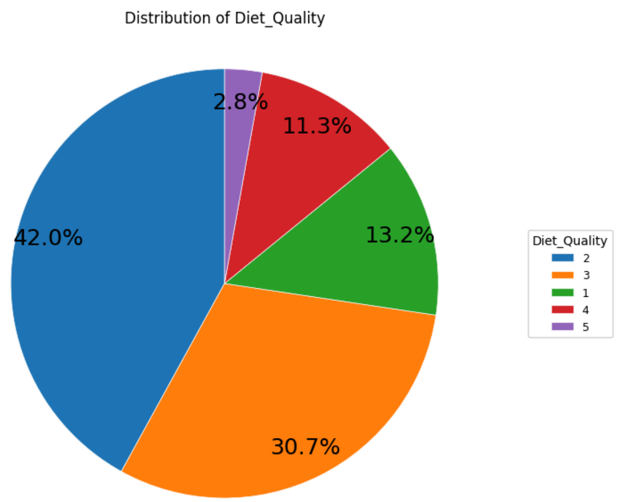


Fig. 3.1(i)

73% of respondents have below-average and average diet quality, while only 14% have a healthy diet.

7. Financial Condition

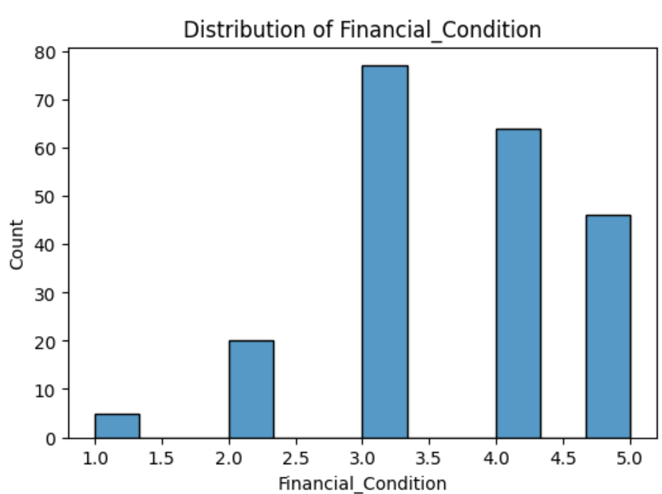


Fig. 3.1(j)

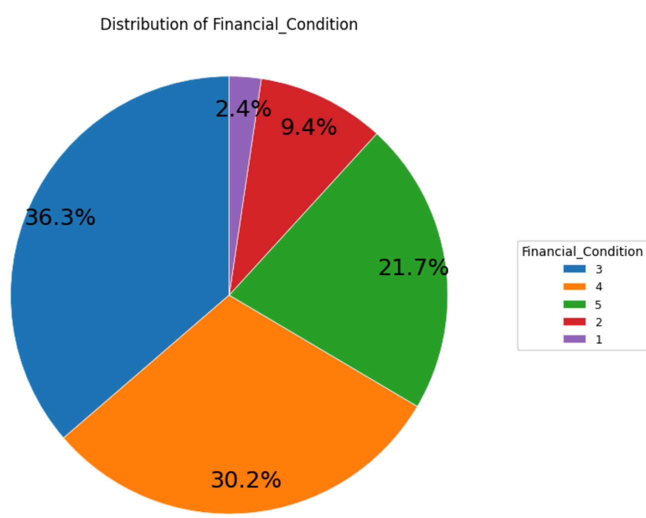
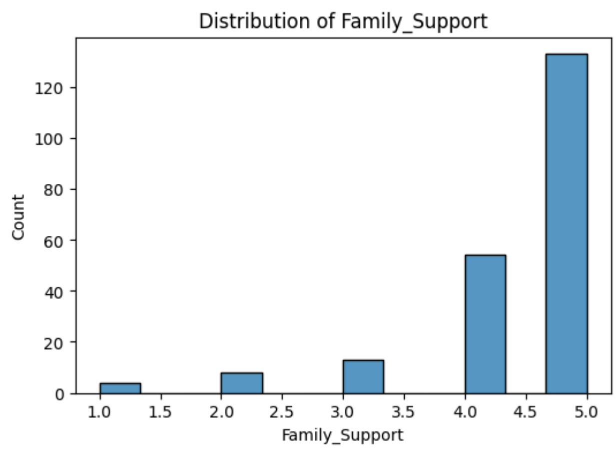


Fig. 3.1(k)

52% of the students are in good financial condition, whereas 12% are in poor financial condition

8. Family Support



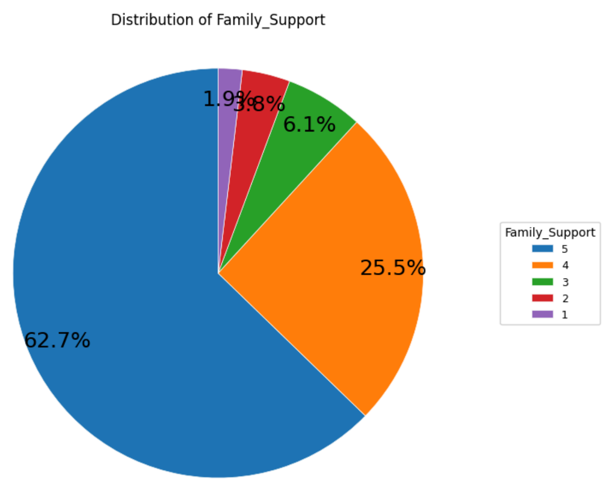


Fig. 3.1(l)

Fig. 3.1(m)

88% of respondents have a supportive family.

4.2.3 Academic Features

1. Side Hustle (Part-Time job)

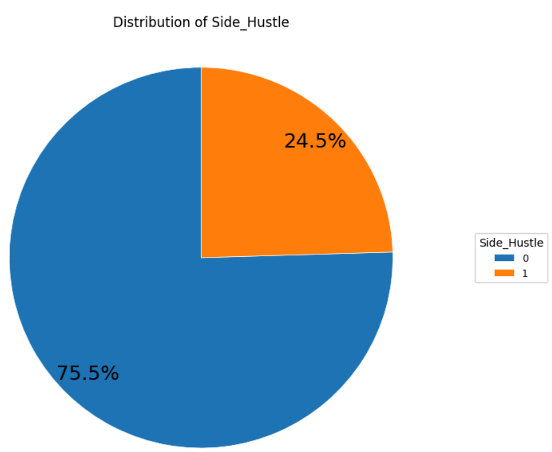


Fig. 3.2(a)

76% of respondents don't have a part-time job, making it a minor factor in predicting stress levels.

2. Study Load

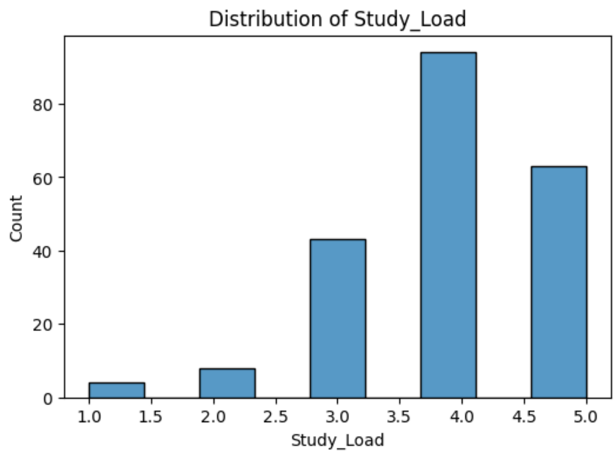


Fig. 3.2(b)

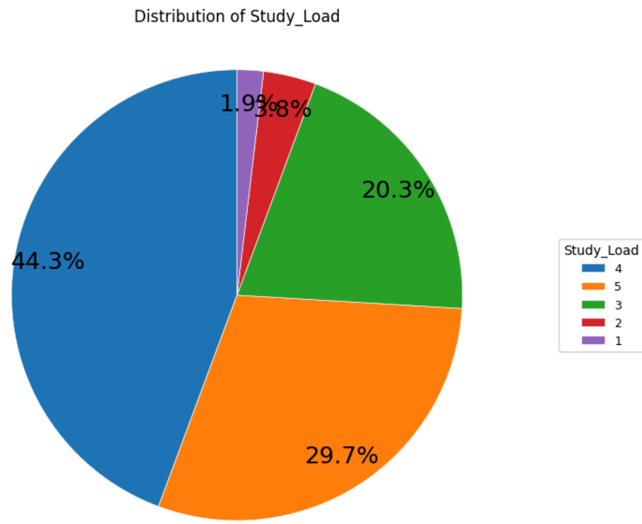


Fig. 3.2(c)

The distribution shows that about 94% have a moderate (3) to high (5) study load.

3. Satisfaction Level with Instructors

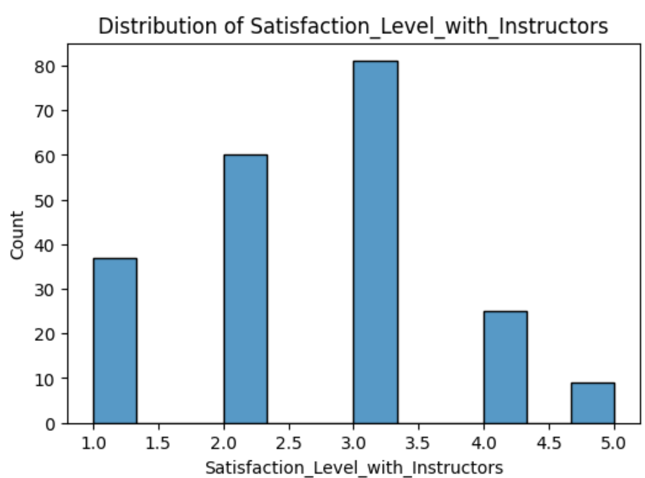


Fig. 3.2(d)

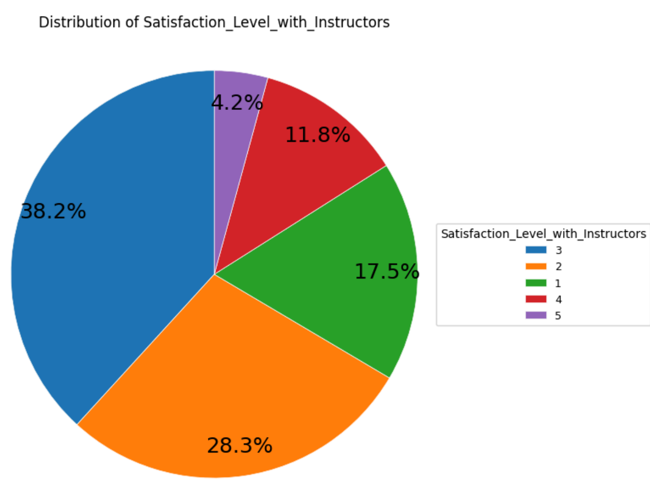


Fig. 3.2(e)

The distribution reveals that 54% of the students are satisfied (3-5) with their instructors, and 46% are not (1-2).

4. Career Concerns

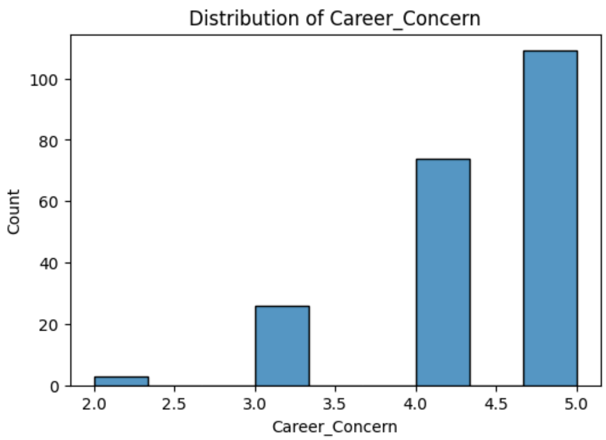


Fig. 3.2(f)

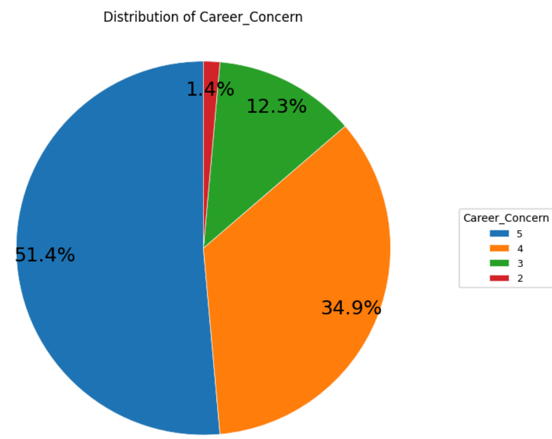


Fig. 3.2(g)

More than 87% of the students are highly concerned about their careers.

5. Peer Pressure

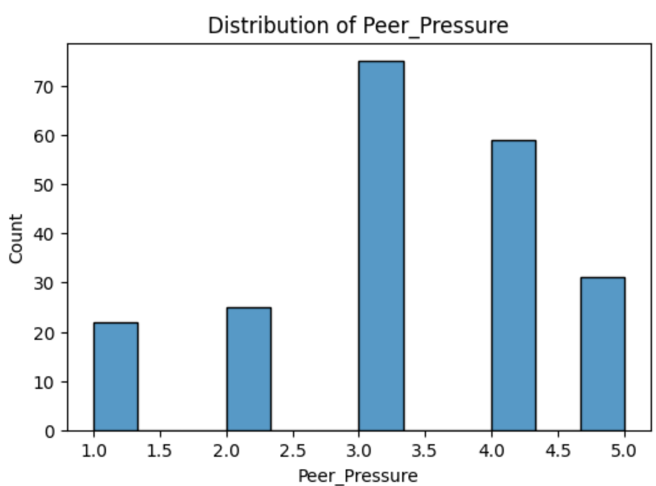


Fig. 3.2(h)

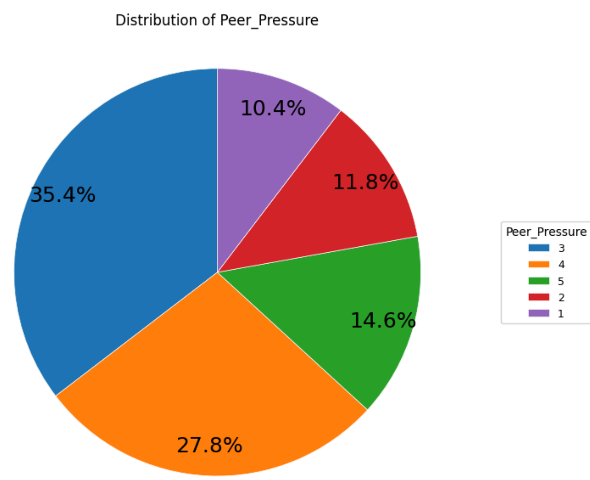
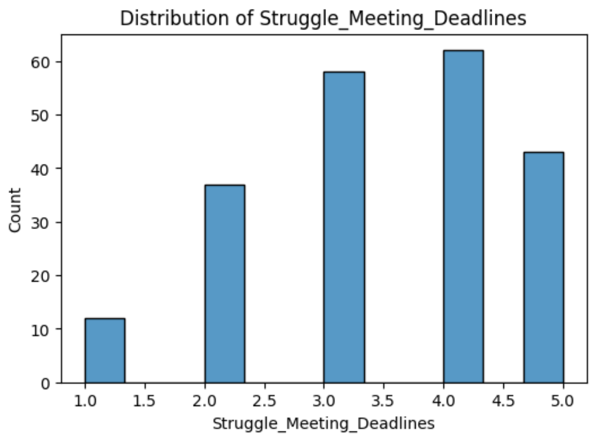


Fig. 3.2(i)

Most people face moderate peer pressure (35.4%), while 27.8% experience high pressure. With fewer reporting low pressure (25%).

6. Struggle Meeting Deadlines



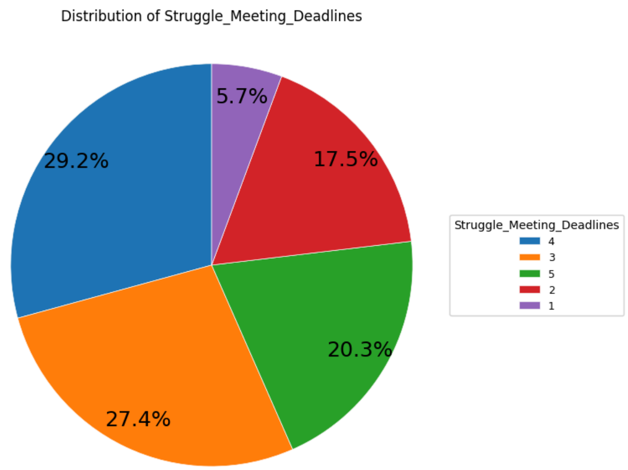
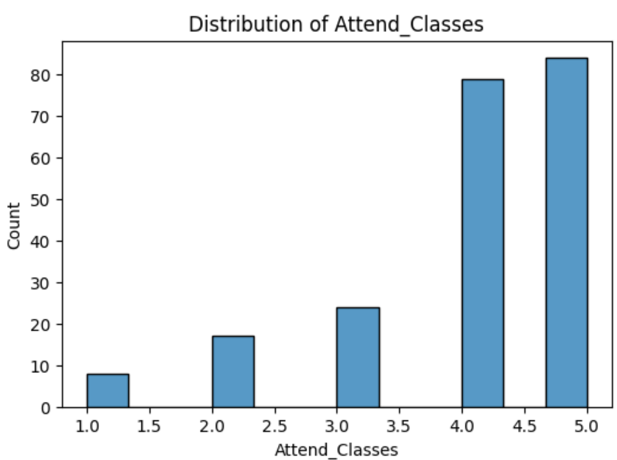


Fig. 3.2(j)

Fig. 3.2(k)

77% of respondents struggle to meet deadlines, indicating significant challenges in time management.

7. Attend Classes



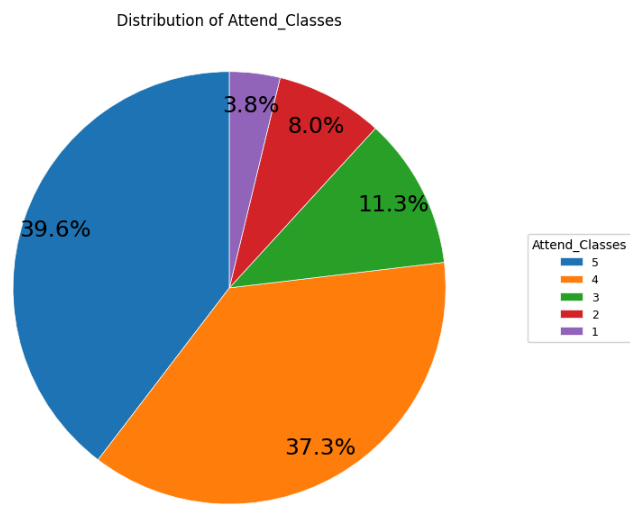


Fig. 3.2(l)

Fig. 3.2(m)

88% of respondents attend classes regularly.

8. Academic Satisfaction

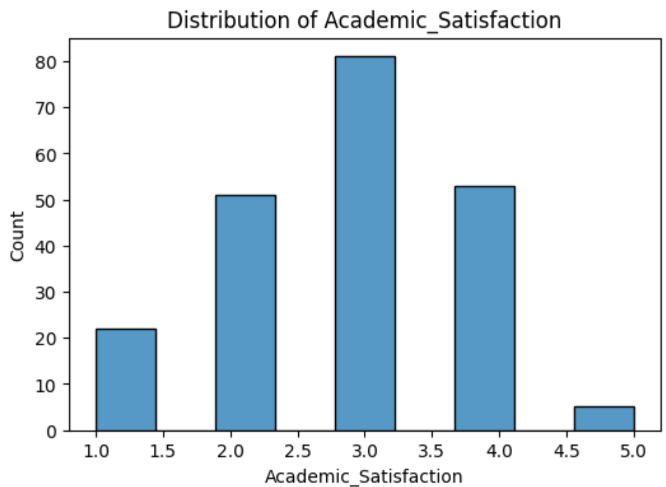


Fig. 3.2(m)

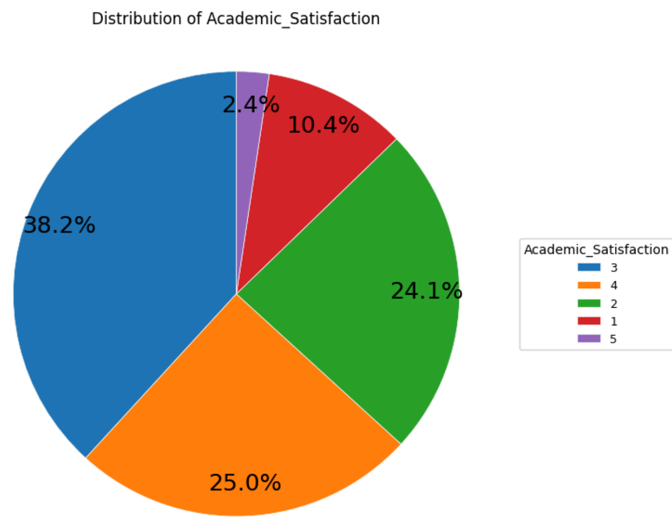


Fig. 3.2(o)

Only 28% of respondents feel academically satisfied.

9. Study Hours

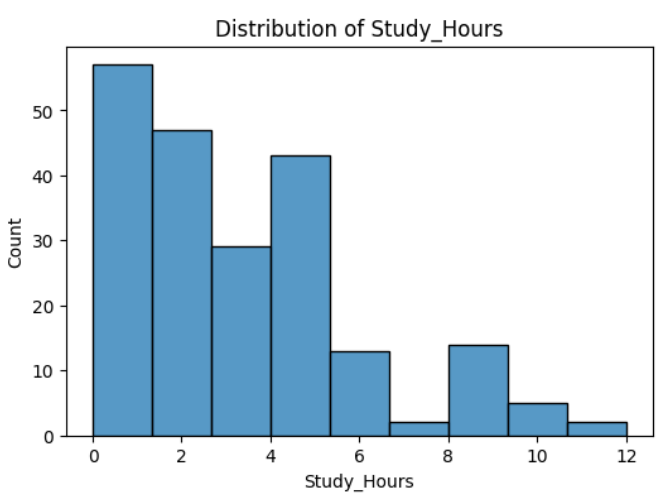


Fig. 3.2(p)

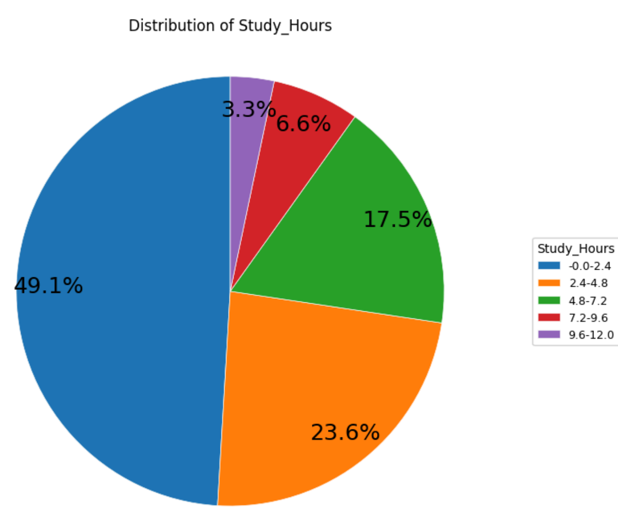


Fig. 3.2(q)

Most individuals (49.1%) study < 2.4 hours, indicating generally low study hours, with fewer devoting extensive time.

10. CGPA

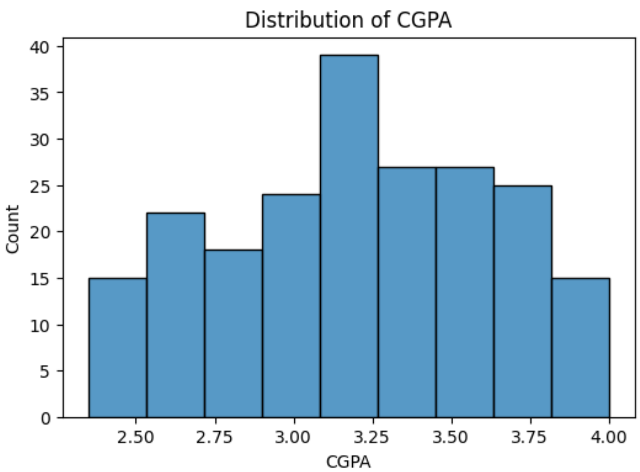


Fig. 3.2(r)

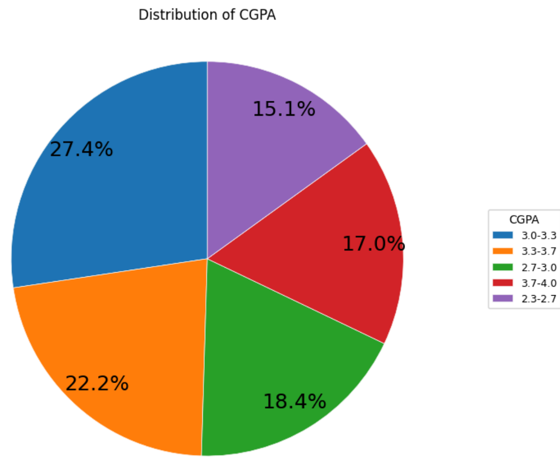


Fig. 3.2(s)

The CGPA distribution follows a bell curve, with most respondents (66.6%) in the moderate range (2.7-3.7). About 18.4% achieve high scores (3.7-4.0), while 15.1% struggle (2.3-2.7).

4.2.4 Personal Factors

1. Mental Health History

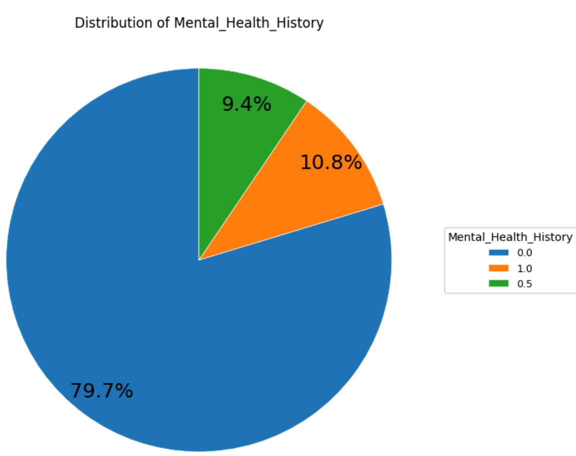
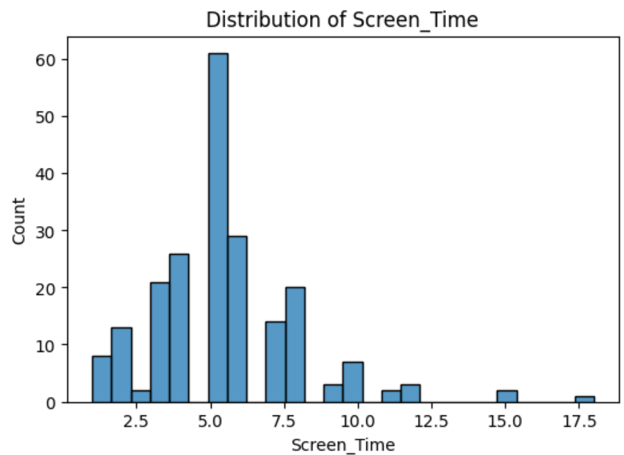


Fig. 3.3(a)

Only a small percentage (11%) of respondents have a history of mental health (1).

2. Screen Time



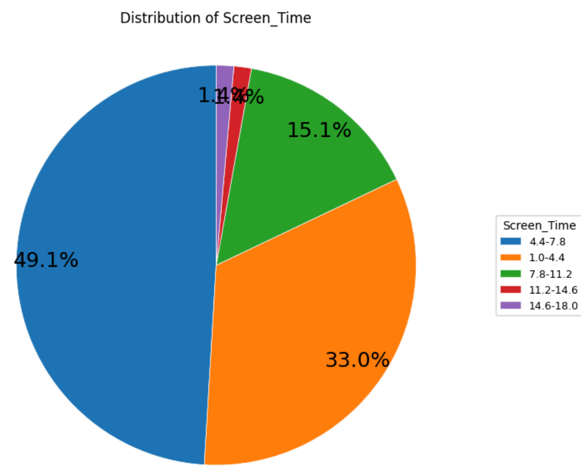
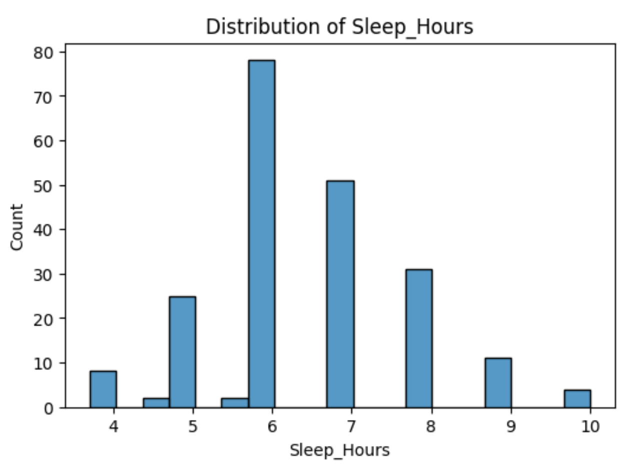


Fig. 3.3(b)

Fig. 3.3(c)

Approximately half (49.1%) of the students have moderate (4.4-7.8) screen time, whereas 18%, i.e., 1 in 5 students, have a screen time of >8 hours.

3. Sleep Hours



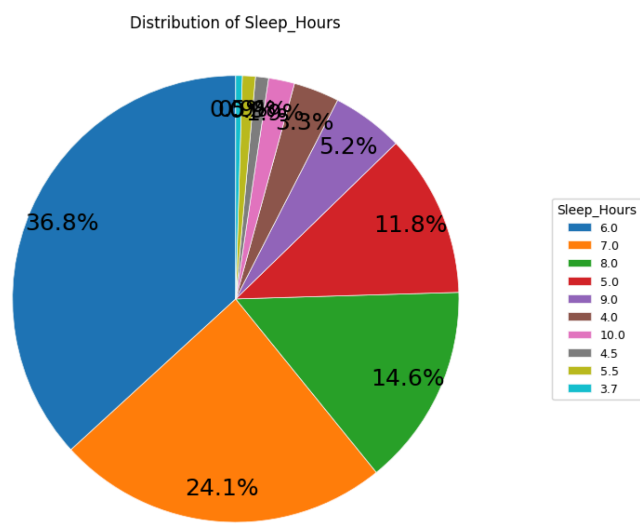
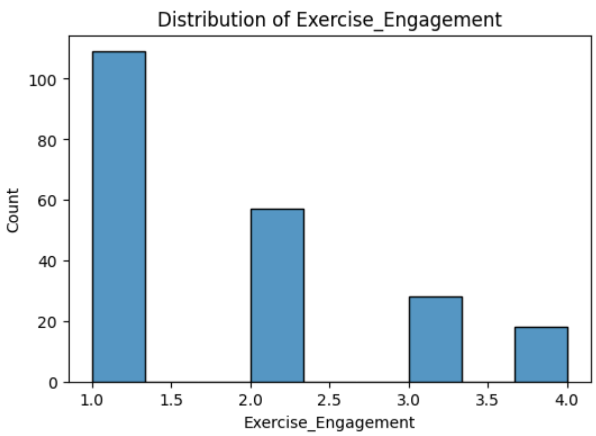


Fig. 3.3(b)

Fig. 3.3(c)

76% of students sleep 6-8 hours, while only 18% sleep for less than 5 hours.

4. Exercise Engagement



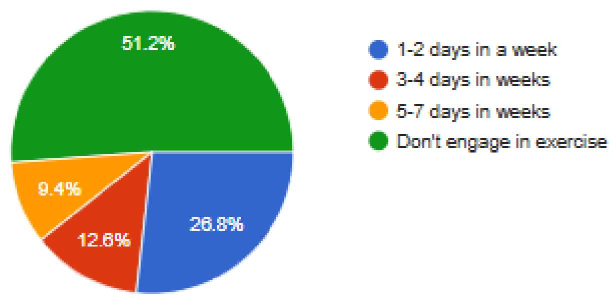
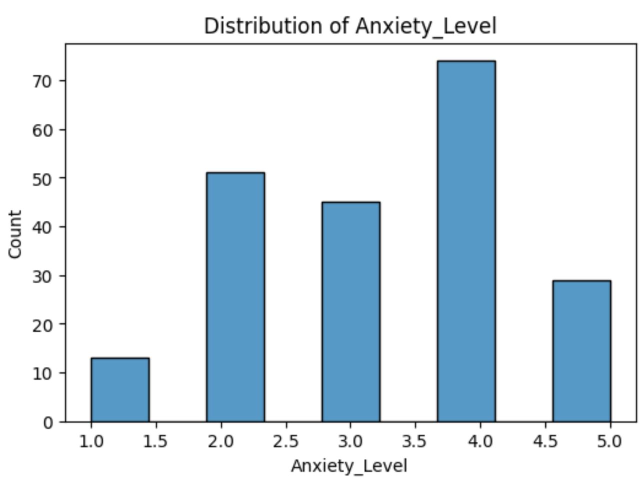


Fig. 3.3(d)

Fig. 3.3(e)

51% of students do not exercise, while only 23 % actively work out throughout the week.

5. Anxiety Level



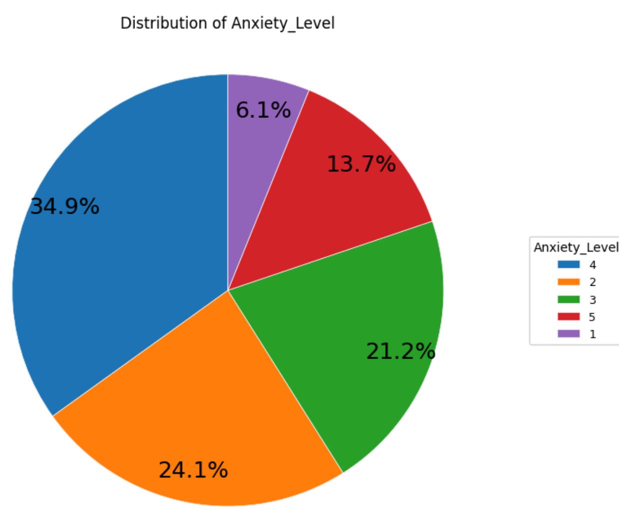
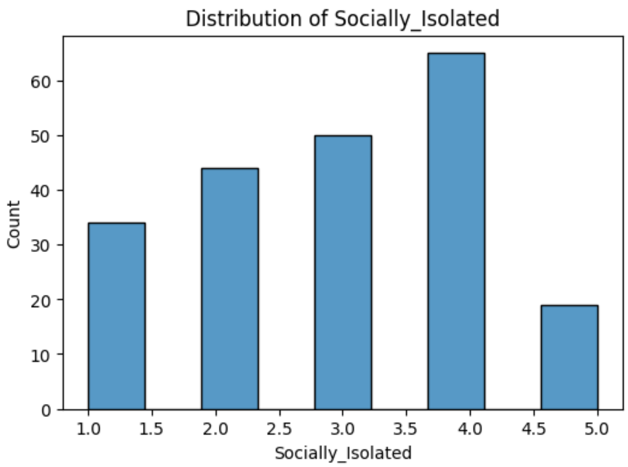


Fig. 3.3(f)

Fig. 3.3(g)

70% of the students have moderate to high anxiety levels.

6. **Socially Isolated**



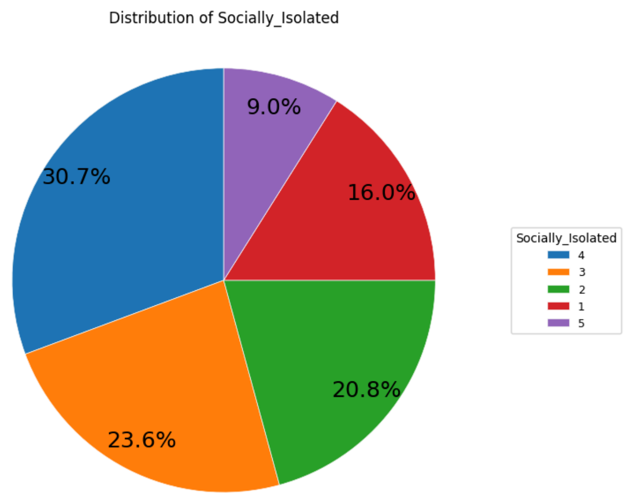
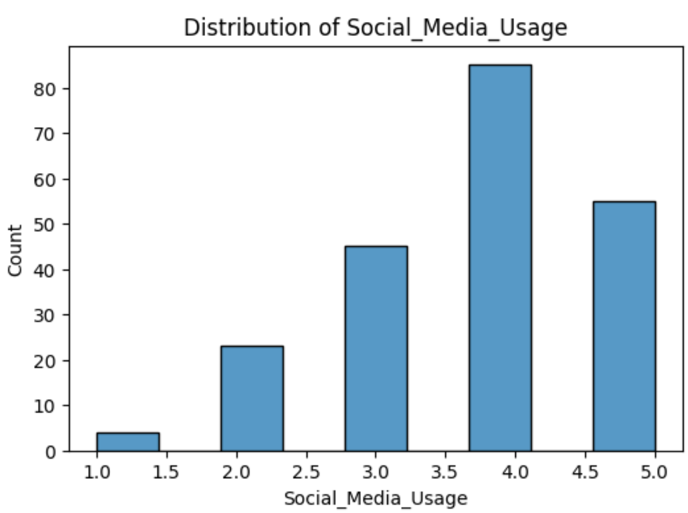


Fig. 3.3(h)

Fig. 3.3(i)

More than 40% of students feel socially isolated or lonely most of the time.

7. **Social Media Usage**



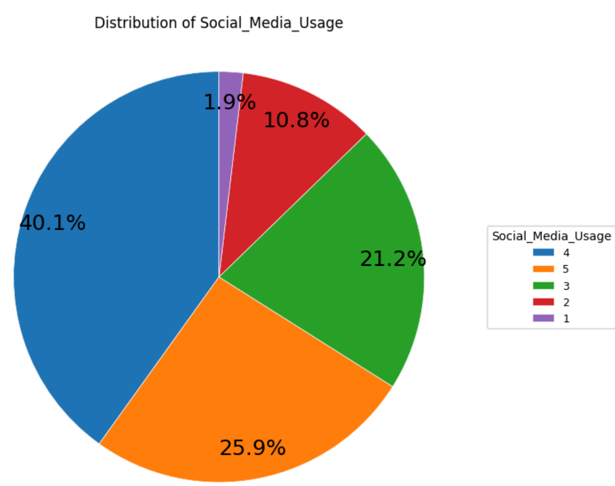
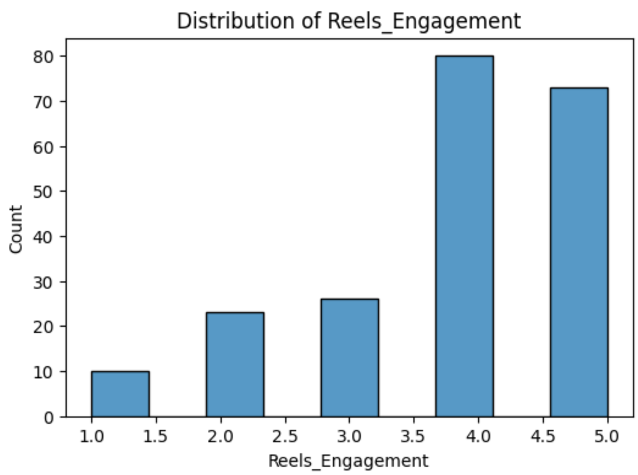


Fig. 3.3(j)

Fig. 3.3(k)

66% of students extensively use social media throughout the day, while 22% use it moderately.

8. Reels Engagement



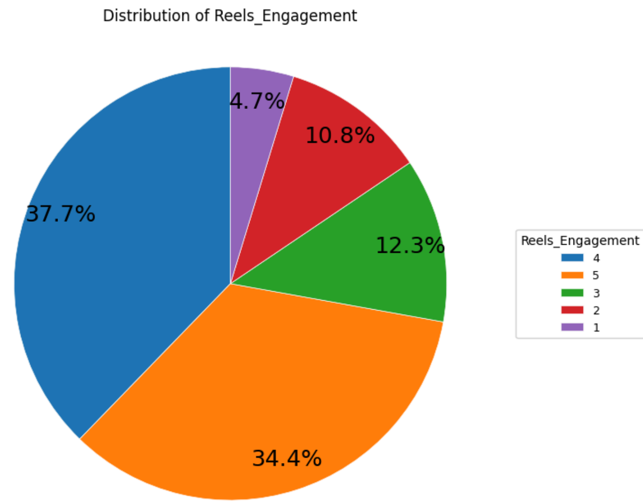
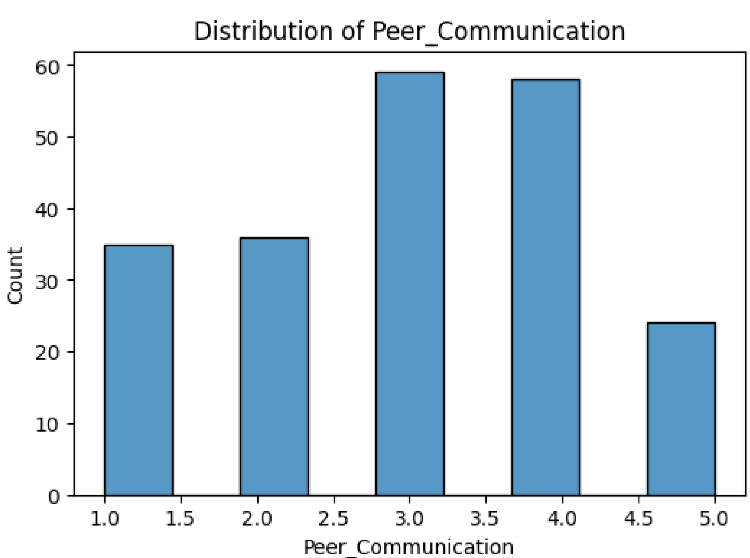


Fig. 3.3(l)

Fig. 3.3(m)

73% of students watch reels frequently throughout the day.

9. Peer Communication



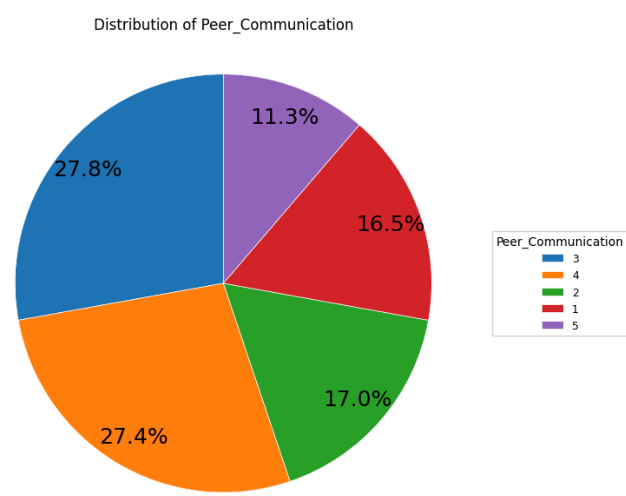


Fig. 3.3(n)

Fig. 3.3(o)

Only 40% of the students communicate with their family or friends when stressed, and 34% avoid this interaction.

10. Peer Communication

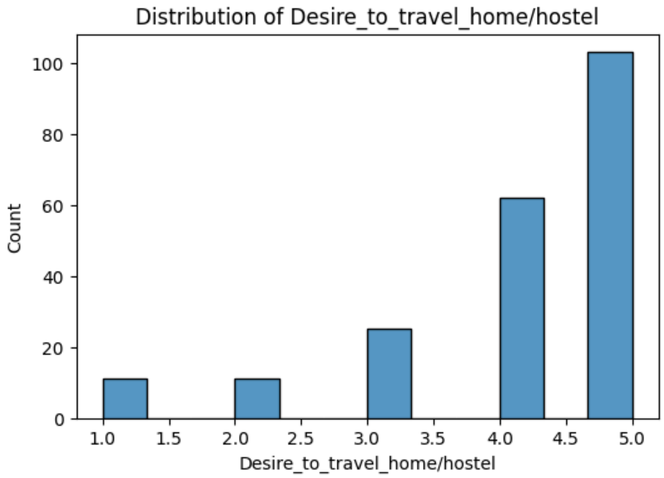


Fig. 3.3(p)

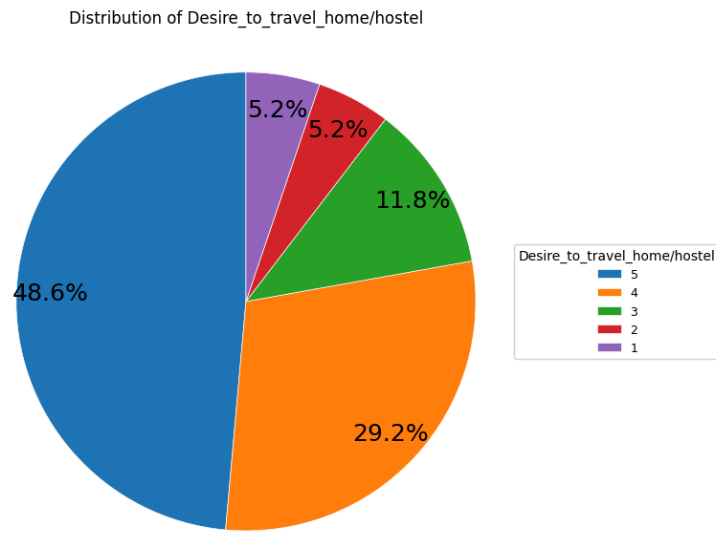


Fig. 3.3(q)

Approximately 70% of the students have a strong desire to go home as soon as classes end.

11. Caffeine Intake

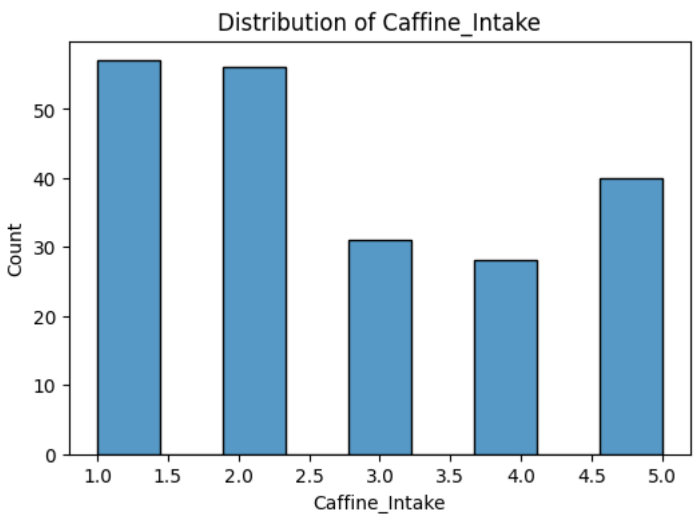


Fig. 3.3(r)

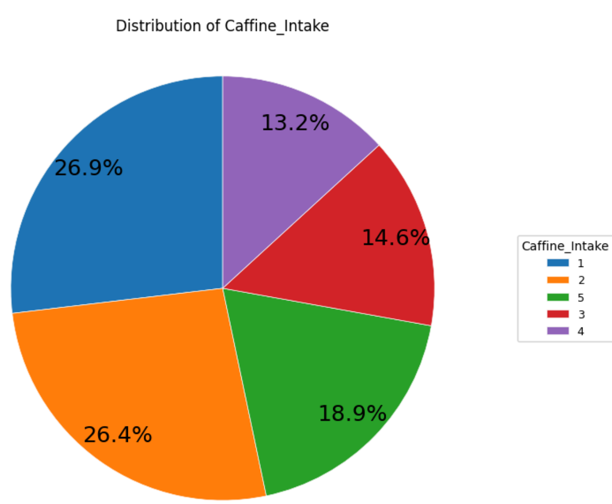


Fig. 3.3(d)

65% of the respondents report little to no caffeine intake, and 32% consume high amounts.

12. Substance Intake

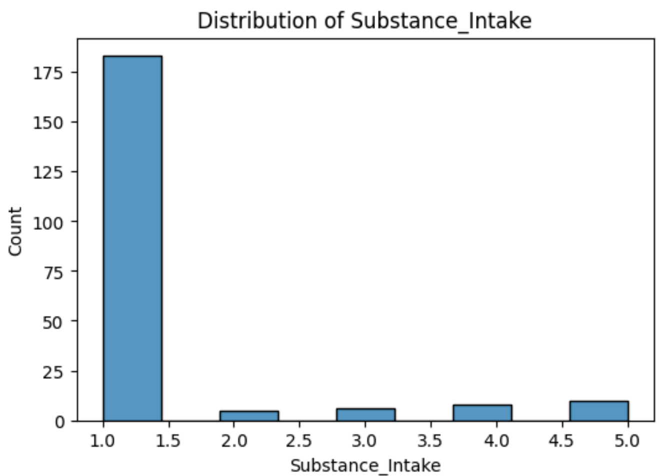


Fig. 3.3(t)

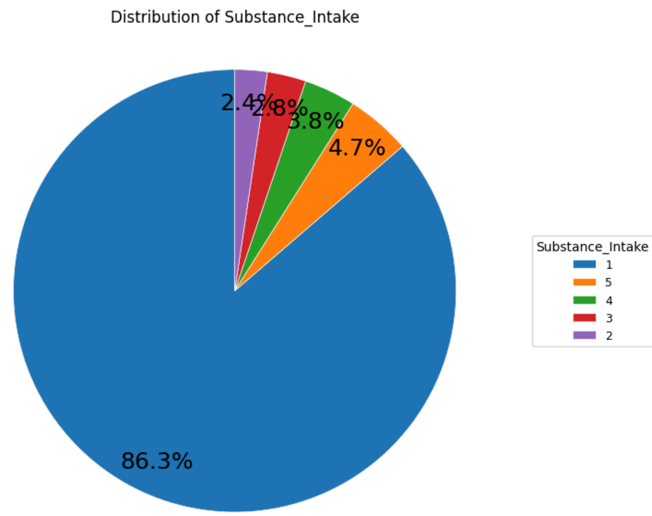


Fig. 3.3(u)

86% of students do not consume any substance, such as cigarettes, to cope with stress.

13. Societies' Participation

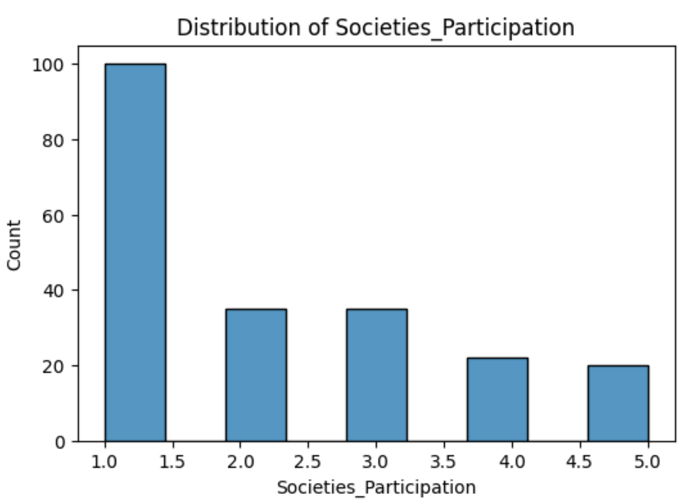


Fig. 3.3(v)

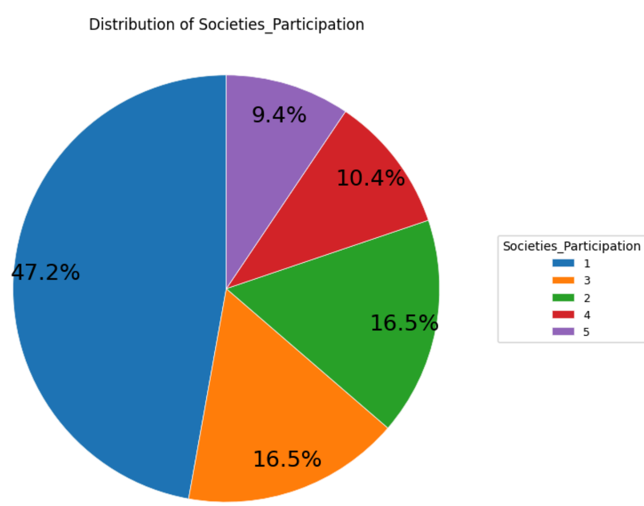


Fig. 3.3(w)

A majority of students (64%) do not engage in clubs or societies, while only 20% participate actively.

14. Social Events Participation

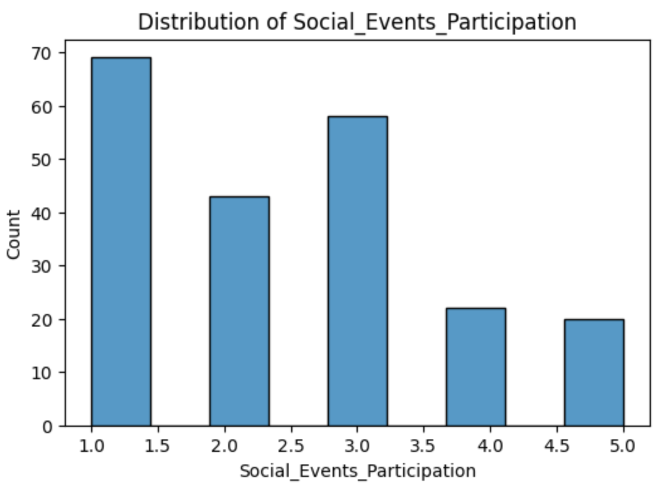


Fig. 3.3(x)

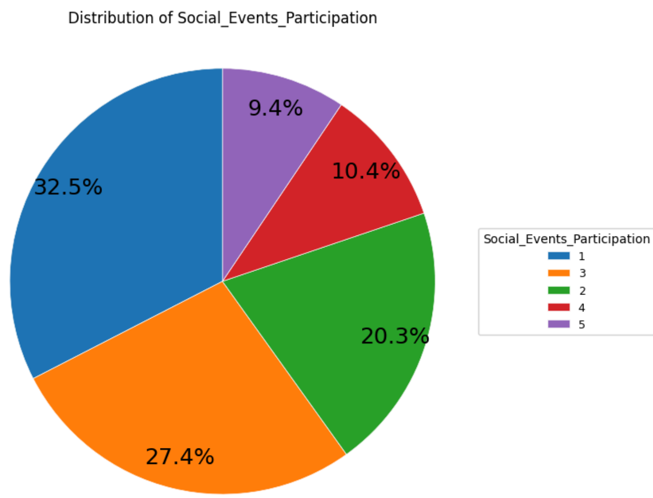


Fig. 3.3(y)

Approximately half of the students (53%) do not attend events, whereas only 11% actively participate.

4.3 Correlation Analysis of Numerical Features

This section focuses on analyzing correlations among the numerical features of the dataset, providing a foundation for interpreting feature importance and predictive power in ML models.

Research Question 3: What demographic, academic, and personal factors have a prominent contribution to stress among students in NUST?

4.3.1 Demographic Features Correlation with Stress_Level

Demographic Features Correlation with Stress_Level:

Feature	Correlation with Stress_Level
Age	0.045
Personality_Type	-0.310
Is_Hostellite	0.187
Diet_Quality	-0.215
Financial_Condition	-0.250
Family_Support	-0.151

Fig. 4(a)

4.3.2 Academic Features Correlation with Stress_Level

Feature	Correlation with Stress_Level
Side_Hustle	-0.224
Study_Load	0.394
Satisfaction_Level_with_Instructors	0.025
Career_Concern	0.059
Peer_Pressure	0.326
Attend_Classes	0.124
Academic_Satisfaction	-0.280
Stress_to_DLM	-0.034
Study_Hours	-0.036
CGPA	0.066

Fig. 4(b)

4.3.3 Personal Features Correlation with Stress_Level

Feature	Correlation with Stress_Level
Mental_Health_History	0.161
Screen_Time	0.090
Sleep_Hours	-0.223
Exercise_Engagement	-0.248
Anxiety_Level	0.745
Socially_Isolated	0.454
Social_Media_Usage	0.271
Reels_Engagement	0.294
Peer_Communication	-0.073
Desire_to_travel_home/hostel	0.126
Caffine_Intake	0.298
Substance_Intake	-0.135
Societies_Participation	0.009
Social_Events_Participation	-0.152

Fig. 4(c)

4.3.4 Top Five Features Correlation with Stress_Level

Feature	Correlation with Stress_Level
Anxiety_Level	0.745
Socially_Isolated	0.454
Study_Load	0.394
Peer_Pressure	0.326
Personality_Type	-0.310

Fig. 4(d)

4.3.5 Correlation (r) Summary with Stress Levels

1. Strong Positive Correlations
   - **Anxiety Level** (r = +0.745): Explains 74.5% of the variance in stress levels, making it the strongest predictor of high stress.
   - **Social Isolation** (r = +0.454): Accounts for 45.4% of stress variance, indicating a strong link between loneliness and elevated stress.
2. Moderate Positive Correlations
   - **Study Load** (r = +0.394): Linked to 39.4% of stress variance; students with heavier academic loads experience more stress.
   - **Peer Pressure** (r = +0.326): Contributes to 32.6% of the variance, showing that social expectations significantly impact stress.
   - **Caffeine Intake** (r = +0.298): Regular or high consumption of caffeine is moderately associated with increased stress levels.
   - **Social Media Usage** (r = +0.271): Excessive use of social media platforms contributes moderately to higher stress.
3. Negative Correlations
   - **Personality Type** (r = -0.310): More extroverted and emotionally stable individuals are less likely to feel stressed.

- **Academic Satisfaction** (r = -0.280): Students who are satisfied with their academic experience tend to experience lower stress.
- **Exercise Engagement** (r = -0.248): Engaging in physical activity helps in reducing stress levels.
- **Financial Condition** (r = -0.250): A stable or strong financial background contributes to lower stress.
- **Diet Quality** (r = -0.215): Better dietary habits are associated with reduced stress.
- **Sleep Quality** (r = -0.223): Good and adequate sleep is linked to lower levels of stress.
- **Family Support** (r = -0.151): The presence of supportive family relationships slightly reduces stress.

4. Weak or No Correlations

- **Age** (r = +0.045): Shows a very weak correlation, indicating age has minimal impact on stress levels.
- **Satisfaction with Instructors** (r = +0.025): Displays a negligible relationship with stress, suggesting instructor satisfaction has little influence.

4.3.6 HeatMap

This is the visualization of the correlations among each attribute of the dataset.

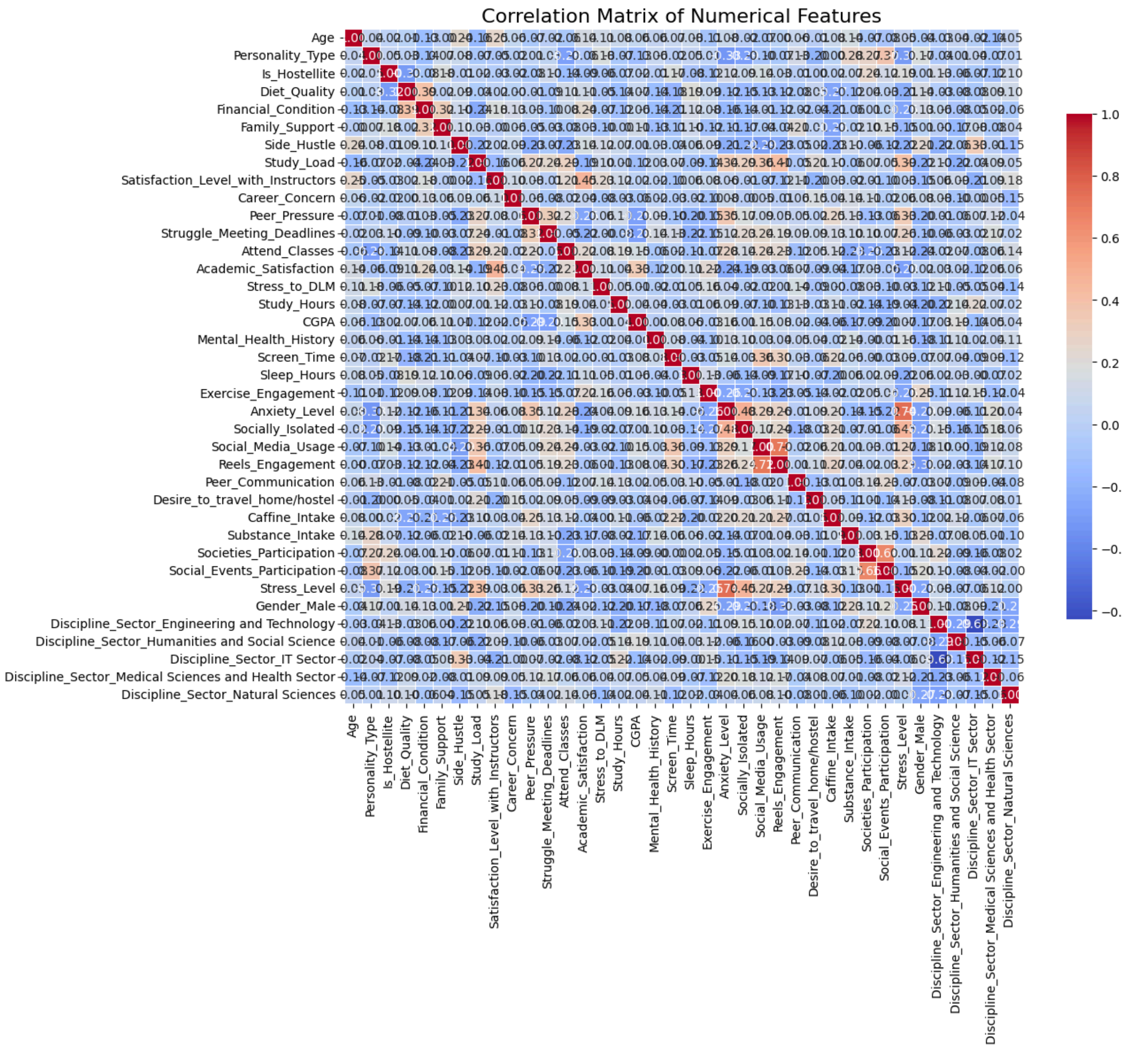


Fig. 4(e)

4.4 Analysis and Evaluation of ML Models Implemented

This section evaluates the efficiency of various ML models in predicting student stress levels using survey data. By analyzing performance metrics, we aim to determine the accuracy and reliability of these algorithms in identifying patterns and consistently forecasting stress levels, highlighting their practical relevance for early detection among students.

Research Question 1: How accurately can machine learning models predict student stress levels based on survey-based data?

Model	Accuracy
Logistic Regression with SMOTE	97.297297
Random Forest without SMOTE	89.189189
SVM with SMOTE	89.189189
XGBoost with SMOTE	86.486486
KNN without SMOTE	83.783784

Table 1.2

The range of accuracies (83.78%–97.30%) demonstrates that machine learning models can predict student stress levels with high accuracy.

- Of the models assessed, Logistic Regression with SMOTE had the *best accuracy* of 97.30%, suggesting that balancing the dataset greatly improves model performance.
- Both with 89.19% accuracy, Random Forest and SVM also turned out to be *consistent classifiers.*
- Although KNN (83.78%) and XGBoost (86.49%) performed somewhat worse, they still indicate a *great predictive power.*

Research Question 2: Which ML algorithm performs best in terms of accuracy, precision, recall, and F1 score in this context?

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression with SMOTE	97.297297	0.975	0.973	0.973
Random Forest without SMOTE	89.189189	0.905	0.865	0.866
SVM with SMOTE	89.189189	0.919	0892	0.894
XGBoost with SMOTE	86.486486	0.865	0.865	0.865
KNN without SMOTE	83.783784	0.892	0.838	0.836

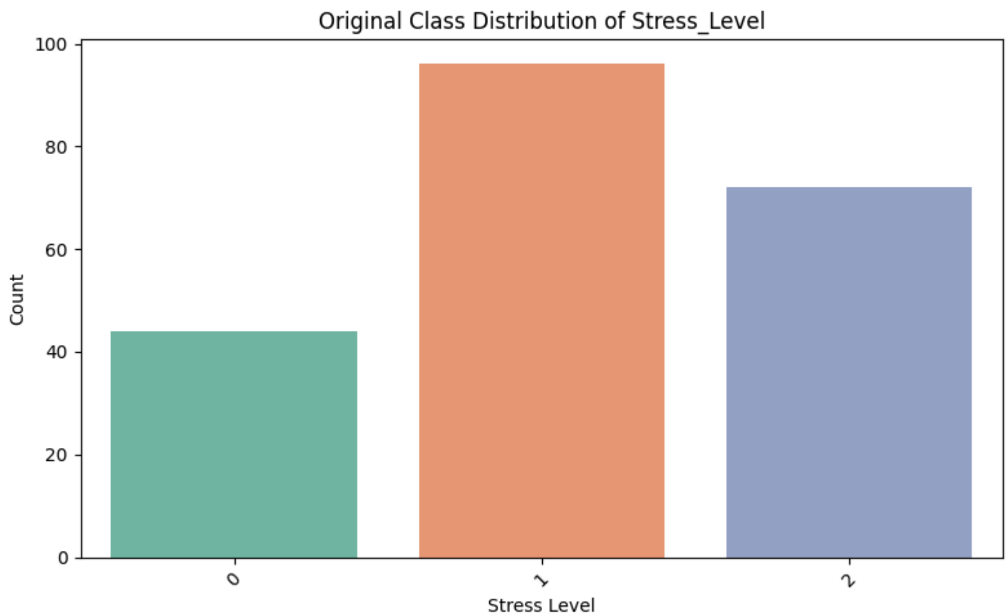
Table 1.3

- Based on all the key evaluation metrics - *accuracy, precision, recall, and F1-score*—Logistic Regression with SMOTE emerged as the most effective model for predicting student stress levels.
- It achieved the highest accuracy (97.3%) and therefore predicted the stress levels for the majority of the cases.
- High precision and recall values indicate that the model not only made a few false predictions but also captured most stress cases successfully.
- Thus, these values show that LR with SMOTE is the most reliable and well-rounded model for this classification task.

Chapter 5: Discussion

5.1 Overview of Stress Level Distribution

The ***first objective*** was to *train* ML models on data collected from university students and to classify them as low, medium, or high stress. Of the 212 instances, 95 were classified as having a medium degree of stress, 71 as having a high level, and 45 as having a low level. These findings suggest that a wide range of students experience moderate to high levels of stress, highlighting a widespread concern among the population.



5.2 Model Performance Analysis

The ***second objective*** of the study was to compare the performance of different machine learning models in predicting stress. So, based on data and findings, we observed that, out of all the models, which included *Random Forests*, *SVM*, *XGBoost*, *KNN*, and *Logistic Regression*, Logistic Regression outperformed the others, achieving an accuracy of 97.29.

5.2.1 Logistic Regression

When paired with SMOTE, logistic regression achieved perfect precision and recall scores for class 2, while classes 0 and 1 had nearly perfect ones. This indicated that logistic regression is highly effective in generalizing patterns in the datasets.

5.2.2 Random Forest without SMOTE

Random Forest, even without SMOTE, performed well for all three classes, achieving almost equal precision and recall as logistic regression with an overall accuracy of *91.89%,* which shows that ensemble models are a really good approach due to their robustness and ability to handle complex data structures.

5.2.3 SVM with SMOTE

The overall accuracy achieved by SVM was 89.19%, and almost a perfect recall score for class 1. The model struggling with class 2 indicates that the model's performance drops with overlapping features.

5.2.4 XGBoost SMOTE (86.49%)

Although XGBoost did well overall, it was only marginally better than Random Forest and SVM. It maintained balanced metrics while performing poorly in any class, which would indicate that XGBoost overfitted a little.

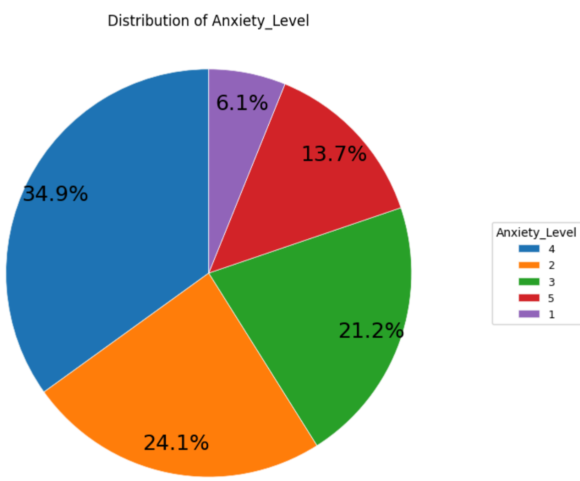
5.2.5 KNN (83.78%)

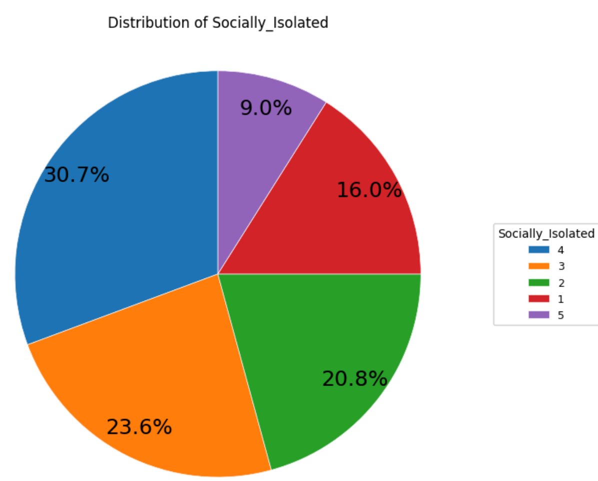
An overall recall of 58% for class 2 indicates that KNN was not a good model. Although it performed well for the high-stress group, the model was generally not reliable.

```
Accuracy Summary
                            Model  Accuracy (%)
Logistic Regression with SMOTE     97.297297
   Random Forest without SMOTE     91.891892
                 SVM with SMOTE     89.189189
             XGBoost with SMOTE     86.486486
              KNN without SMOTE     83.783784
```

5.3 Influence of Demographic and Academic Factors

The *third objective* was to identify the most influential factors affecting the stress among students. *Correlation Analysis* suggested that social isolation and anxiety strongly impact stress. Students' stress levels were moderately affected by their study load, peer pressure, caffeine intake, and social media engagement. Conversely, stress is generally lessened by extroverted personalities, academic satisfaction, physical activity, adequate sleep, and a healthy diet. There was almost no correlation between stress and either age or teacher satisfaction.





In a nutshell, it is clear from the study that a large number of students face medium to high levels of stress influenced by anxiety and academic workload. There is a dire need to address the rising stress among students. Educational institutions must prioritize the development of mental health support systems, helping students cope with stress.

Chapter 06: Conclusion

The study reveals moderate to high levels of stress among students, highlighting that there is a need to address the mental well-being of students. The finding also identifies that factors such as social isolation, anxiety, and academic pressure play a significant role in shaping stress, while healthy habits and academic satisfaction tend to offer protective effects.

Of all the models, LR appeared to be the most effective for our dataset, achieving the best accuracy and balanced performance across all the classes. Random Forest showed reliable results even without data balancing. SVM and XGBoost also performed well, but with limitations in handling certain classes. KNN struggled and was less reliable in this context.

Overall, the analysis not only identifies the key factors affecting stress but also highlights the impact of data handling techniques such as SMOTE in enhancing the performance of ML models.

Chapter 07: Recommendations

7.1 Practical Recommendations (Individual Level)

- **Promote Digital Detox Intervals:**
  - Students should practice "digital well-being" by limiting screen time during non-academic hours. This can be encouraged through awareness campaigns, mobile app timers, and campus challenges.
- **Adopt Personalized Mental Health Tracking:**
  - Encourage students to self-monitor stress with ML-based mobile applications that use behavioral patterns (sleep, screen time) to suggest wellness activities like meditation or breaks.
- **Organize Peer-Support and Counselling Circles:**
  - Peer-led sessions can foster open discussions and emotional resilience. ML predictions can help proactively invite students showing signs of elevated stress.
- **Stress Awareness Workshops:**
  - Universities should hold monthly workshops based on recent ML findings, guiding students in stress management based on their own academic, lifestyle, and emotional data patterns.

7.2 Research-Based Recommendations (Future Work)

- **Incorporate Physiological and Sensor-Based Data:**
  - Future research can expand by integrating wearable data (e.g., heart rate, sleep cycles) alongside survey responses for more accurate stress detection.
- **Expand the Dataset and Demographics:**
  - Include more diverse academic institutions and age groups across different Pakistani provinces to improve model generalizability and regional insights.
- **Multi-Label Classification & Deep Learning:**

    - Use advanced models like LSTM, CNN, or transformers for modeling time-series behavioral data, improving accuracy for early-stage stress detection.
- **Real-Time Monitoring System:**
    - Create integrated university dashboards that allow counselors to monitor aggregated, anonymized stress scores in real-time using live student data (e.g., LMS activity, attendance).

References

Alahmadi, N. S., & Hussain, S. (2022). Machine learning-based predictive modelling for mental health risk assessment in university students. *Journal of Education and Information Technologies, 27*(6), 8773–8787. https://doi.org/10.1007/s10639-022-11200-5

Alghamdi, M., Al-Muallim, A., & Algarny, M. (2022). Detection of stress using machine learning techniques: A survey. *Journal of Ambient Intelligence and Humanized Computing, 13*, 5499–5519. https://doi.org/10.1007/s12652-021-03573-3

Arora, A., & Malik, P. (2022). Comparative analysis of classifiers for stress level prediction among students. *Procedia Computer Science, 198*, 131–138. https://doi.org/10.1016/j.procs.2021.12.198

Bokolo, A. J., & Liu, Y. (2023). Machine learning-based framework for stress prediction among university students. *Computers in Human Behavior Reports, 9*, 100234. https://doi.org/10.1016/j.chbr.2023.100234

Choudhury, T., & Basu, S. (2020). Predicting stress using behavioural traits and supervised learning. *International Journal of Cognitive Informatics and Natural Intelligence, 14*(3), 1–20. https://doi.org/10.4018/IJCINI.2020070101

Kaur, P., & Sharma, M. (2020). Machine learning approaches in mental health prediction: Review and future directions. *Healthcare Technology Letters, 7*(6), 177–181. https://doi.org/10.1049/htl.2020.0012

Muhammad, A., & Habib, M. (2023). AI-powered early stress detection system for college students in Pakistan. *Asian Journal of Psychiatry, 87*, 103688. https://doi.org/10.1016/j.ajp.2023.103688

Nguyen, T., & Nguyen, Q. (2021). Mental health status prediction using survey-based data and machine learning techniques. *Heliyon, 7*(12), e08622. https://doi.org/10.1016/j.heliyon.2021.e08622

Rashid, M., Hussain, I., & Sheikh, A. A. (2021). Predicting psychological stress using machine learning techniques. *International Journal of Mental Health and Addiction*. https://doi.org/10.1007/s11469-021-00564-2

Salehinejad, H., & Beheshti, A. (2019). Deep learning for mental health prediction in educational settings. *IEEE Access, 7*, 22100–22110. https://doi.org/10.1109/ACCESS.2019.2897111

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine, 49*(9), 1426–1448. https://doi.org/10.1017/S0033291719000151

Vaibhav, A., & Vyas, D. (2021). A hybrid model for mental health prediction using machine learning algorithms. *Journal of Medical Systems, 45*(5), 52. https://doi.org/10.1007/s10916-021-01712-9

Yadav, A. K., & Singh, J. P. (2020). Analysing stress among university students using machine learning. *Procedia Computer Science, 167*, 2352–2361. https://doi.org/10.1016/j.procs.2020.03.289

Sharma, R. (2021). Logistic Regression from scratch (Multi-classification with OneVsAll). *Medium*. https://medium.com/analytics-vidhya/logistic-regression-from-scratch-multi-classification-with-onevsall-d5c2acf0c37c

Cano, I., Agarwal, S., Ghandehari, A., & Abbasi, A. (2021). A schematic depiction of the classification flow of Random Forest with three Decision Trees. *ResearchGate*. https://www.researchgate.net/figure/A-schematic-depiction-of-the-classification-flow-of-Random-Forest-with-three-Decision_fig3_343223411

Mori, K., & Ono, Y. (2023). Neural networks for clinical prediction models. In *Deep Learning and Artificial Intelligence in Healthcare* (pp. 139-152). Elsevier. https://www.sciencedirect.com/science/article/abs/pii/B9780443184987000090

Bensalem, F. B., Hamrouni, K., Solaiman, B., & Ellouze, N. (2021). The effect of deep feature concatenation in the classification problem: An approach on COVID-19 disease detection.

ResearchGate.
https://www.researchgate.net/publication/355194669_The_effect_of_deep_feature_concatenation_in_the_classification_problem_An_approach_on_COVID-19_disease_detection

Omarzai, F. (2023). XGBoost classification in depth. *Medium*. https://medium.com/@fraidoonomarzai99/xgboost-classification-in-depth-979f11ef4bf9

Soni, S. (2022). K-Nearest Neighbours — Introduction to Machine Learning Algorithms. *Medium*.
https://medium.com/@sachinsoni600517/k-nearest-neighbours-introduction-to-machine-learning-algorithms-9dbc9d9fb3b2