PREDICTIVE ANALYTICS

PROJECT REPORT

(Project Semester: October - December 2025)

Student Lifestyle & Well-being Survey

Submitted by

Vansh Garg

Registration No.12314181

Programme and Section K23BM

Course Code: INT - 234

Under the Guidance of

Dr. Tanima Thakur (UID:23532)

Discipline of CSE/IT

Lovely School of Computer Science

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that **Vansh Garg**, bearing Registration no. 12314181, has completed the INT 234 project titled **"Student Lifestyle & Well-being Survey**" under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort, and study.

Dr. Tanima Thakur (UID:23532)

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab.

Date: 17-12-2025

DECLARATION

I, Vansh Garg, student of BTech under CSE Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 17-12-2025 Signature: **Vansh Garg**

Registration No. 12314181

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my guide, Dr. Tanima Thakur, for their invaluable guidance, support, and encouragement throughout the completion of this project. His expertise and constructive feedback have greatly contributed to the success of this work.

I would also like to extend my sincere thanks to Lovely Professional University for providing me such a wonderful opportunity to work on this project in the subject Predictive Analytics with subject code INT 234, helping with the necessary resources and skills that laid the foundation for my research.

This project, titled "**Student Lifestyle & Wellbeing Survey** ", has been a learning experience, and I would like to acknowledge the support of my peers, family, and all others who helped me in any manner.

Thank you all for your continuous support and motivation.

Name: Vansh Garg
Section: K23BM

Roll No.: 09

Reg. No.: 12314181
Date: 17-12-2025

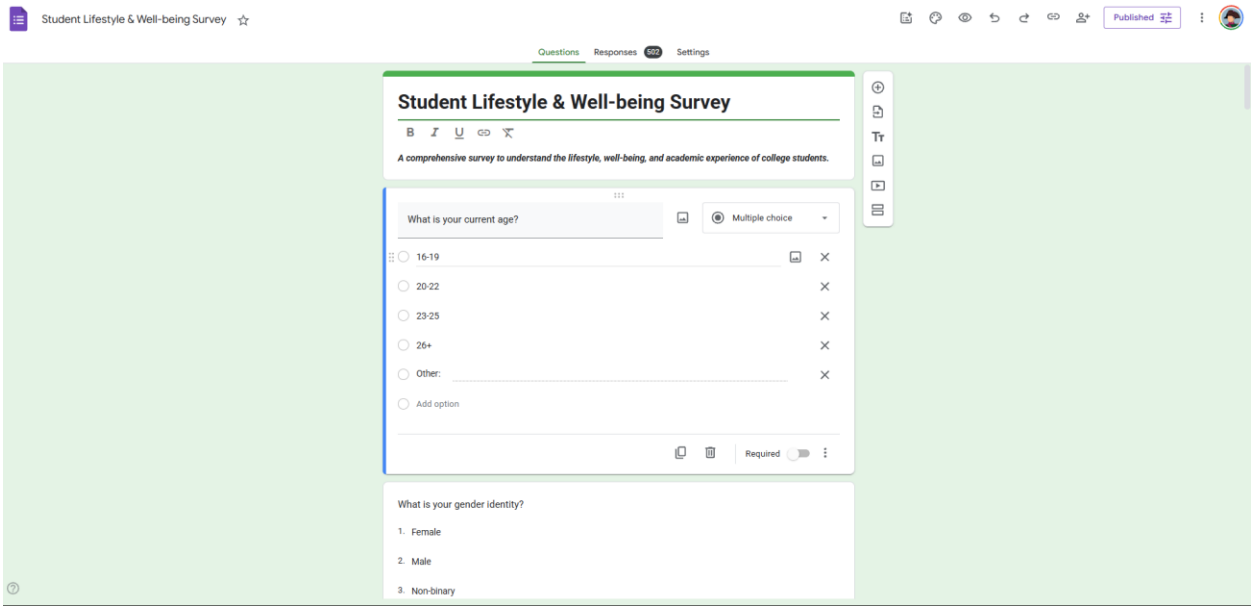
CONTENTS

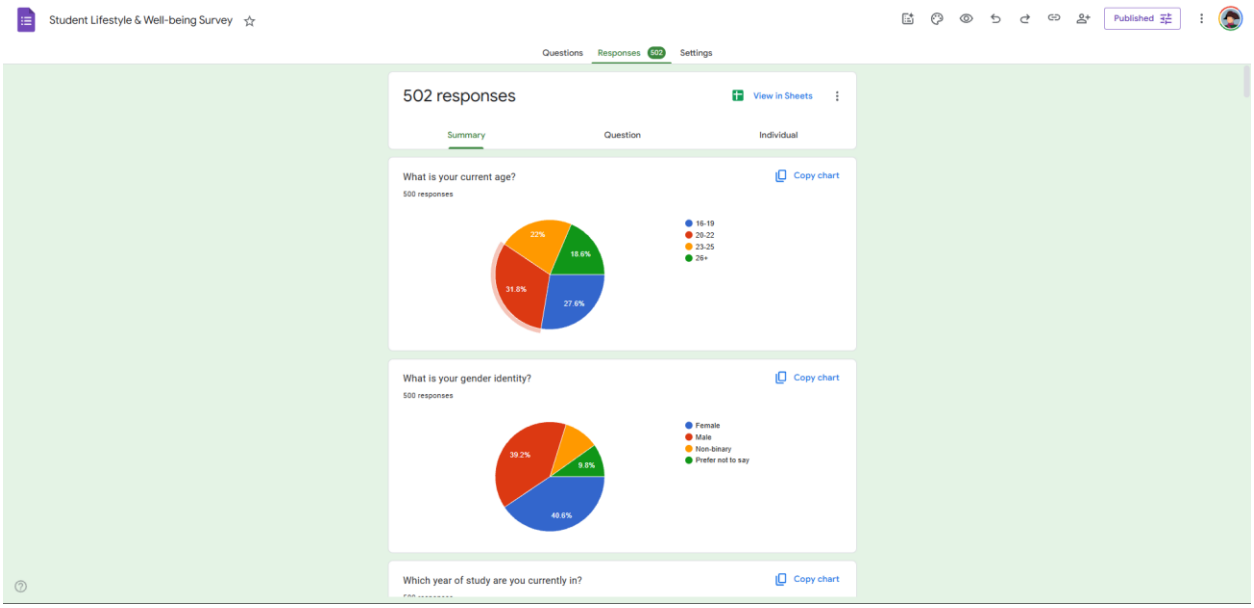
Abstract

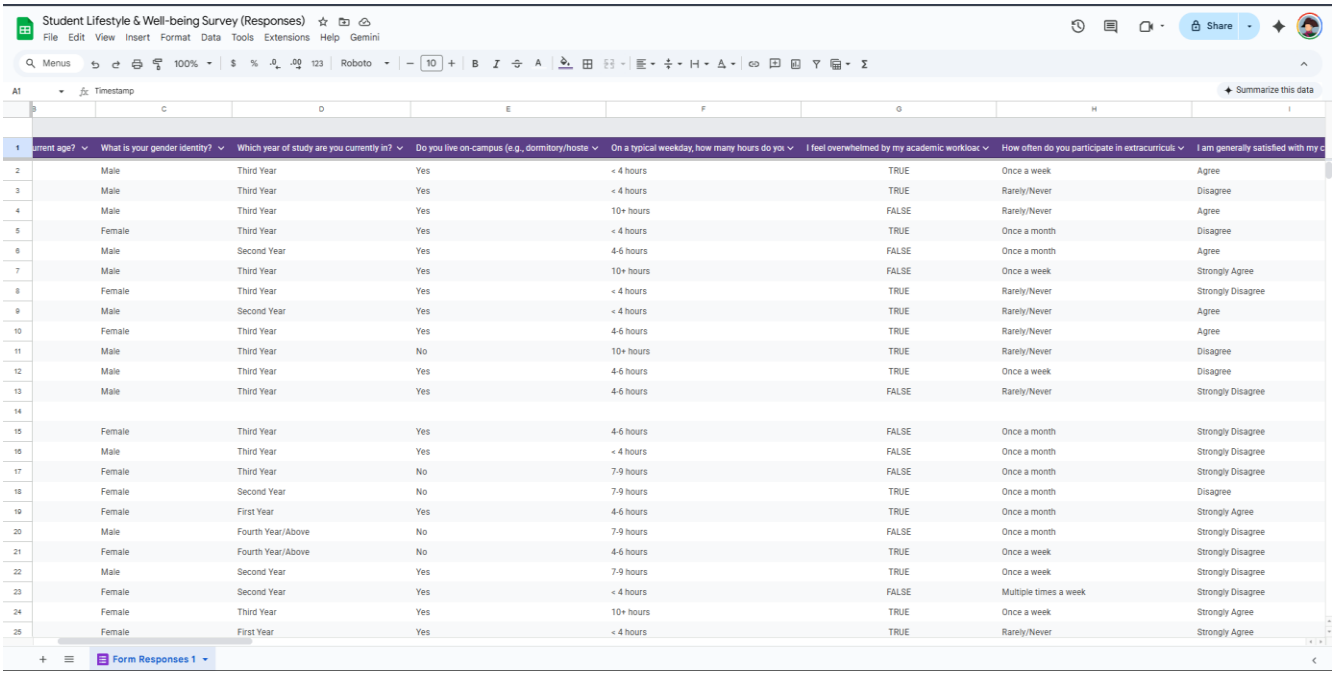
Predictive Analytics has emerged as a powerful tool for extracting meaningful insights from data and supporting informed decision-making across various domains. In the academic environment, students often face challenges related to stress, lifestyle imbalance, academic workload, and mental well-being. This project presents a comprehensive predictive analytics study aimed at understanding student lifestyle patterns and stress management behavior using a self-collected survey dataset. The dataset was created by the author using Google Forms and consists of 500 student responses with 41 distinct questions covering areas such as sleep habits, academic pressure, daily routines, mental health, and stress perception.

The study follows a structured data science workflow beginning with data preprocessing, including removal of irrelevant attributes, handling of missing values, normalization of categorical responses, and encoding of Likert-scale variables. Exploratory Data Analysis (EDA) was conducted to examine response distributions and identify key trends within the dataset. Correlation analysis was performed to investigate relationships between lifestyle factors and stress management. Supervised learning techniques such as Multiple Linear Regression and Logistic Regression were applied to model stress-related outcomes, while model performance was evaluated using metrics including MAE, MSE, RMSE, R^2 score, accuracy, precision, recall, and F1-score. Unsupervised learning techniques such as K-Means clustering and Principal Component Analysis (PCA) were used to group students based on behavioral similarities and visualize high-dimensional data.

The results of this study demonstrate that lifestyle factors have a measurable influence on students' ability to manage stress. The project highlights the effectiveness of predictive analytics in analyzing survey-based data and provides valuable insights that can support academic institutions in designing data-driven student well-being initiatives.







	urrent age?	What is your gender identity?	Which year of study are you currently in?	Do you live on-campus (e.g., dormitory/hoste	On a typical weekday, how many hours do you	I feel overwhelmed by my academic workloac	How often do you participate in extracurricula	I am generally satisfied with my c
2		Male	Third Year	Yes	< 4 hours	TRUE	Once a week	Agree
3		Male	Third Year	Yes	< 4 hours	TRUE	Rarely/Never	Disagree
4		Male	Third Year	Yes	10+ hours	FALSE	Rarely/Never	Agree
5		Female	Third Year	Yes	< 4 hours	TRUE	Once a month	Disagree
6		Male	Second Year	Yes	4-6 hours	FALSE	Once a month	Agree
7		Male	Third Year	Yes	10+ hours	FALSE	Once a week	Strongly Agree
8		Female	Third Year	Yes	< 4 hours	TRUE	Rarely/Never	Strongly Disagree
9		Male	Second Year	Yes	< 4 hours	TRUE	Rarely/Never	Agree
10		Female	Third Year	Yes	4-6 hours	TRUE	Rarely/Never	Agree
11		Male	Third Year	No	10+ hours	TRUE	Rarely/Never	Disagree
12		Male	Third Year	Yes	4-6 hours	TRUE	Once a week	Disagree
13		Male	Third Year	Yes	4-6 hours	FALSE	Rarely/Never	Strongly Disagree
14								
15		Female	Third Year	Yes	4-6 hours	FALSE	Once a month	Strongly Disagree
16		Male	Third Year	Yes	< 4 hours	FALSE	Once a month	Strongly Disagree
17		Female	Third Year	No	7-9 hours	FALSE	Once a month	Strongly Disagree
18		Female	Second Year	No	7-9 hours	TRUE	Once a month	Disagree
19		Female	First Year	Yes	4-6 hours	TRUE	Once a month	Strongly Agree
20		Male	Fourth Year/Above	No	7-9 hours	FALSE	Once a month	Strongly Disagree
21		Female	Fourth Year/Above	No	4-6 hours	TRUE	Once a week	Strongly Disagree
22		Male	Second Year	Yes	7-9 hours	TRUE	Once a week	Strongly Disagree
23		Female	Second Year	Yes	< 4 hours	FALSE	Multiple times a week	Strongly Disagree
24		Female	Third Year	Yes	10+ hours	TRUE	Once a week	Strongly Agree
25		Female	First Year	Yes	< 4 hours	TRUE	Rarely/Never	Strongly Agree

Introduction

In recent years, the rapid growth of data generation has led to an increased demand for advanced analytical techniques capable of extracting meaningful insights from large and complex datasets. Predictive Analytics, a key domain within data science and machine learning, focuses on analyzing historical and current data to identify patterns, relationships, and trends that can be used to make informed predictions about future outcomes. Its applications span multiple fields, including healthcare, finance, marketing, education, and social sciences. In the context of education, predictive analytics plays a crucial role in understanding student behavior, academic performance, and mental well-being.

Student life is often characterized by academic pressure, time constraints, lifestyle imbalance, and psychological stress. Factors such as irregular sleep patterns, excessive workload, poor time management, and lack of physical activity can significantly affect a student's mental health and overall well-being. Stress, if not managed effectively, may lead to decreased academic performance, burnout, and long-term health issues. Therefore, it is essential to analyze student lifestyle and stress-related factors using data-driven approaches that can provide objective insights and support timely interventions.

This project aims to analyze student lifestyle and stress management patterns using predictive analytics techniques applied to a real-world survey dataset. The dataset used in this study was self-collected by the author through a Google Form designed to capture various aspects of student life, including daily routines, academic workload, sleep habits, and stress perception. The survey responses provide a rich source of categorical data, primarily based on Likert-scale questions, making it suitable for exploratory data analysis, correlation studies, and both supervised and unsupervised learning techniques.

The primary objective of this project is not only to gain insights into student stress and lifestyle behaviors but also to demonstrate the practical implementation of predictive analytics concepts taught in the INT234 course. The project follows a systematic analytical pipeline that includes data preprocessing, exploratory data analysis, model building using regression and classification algorithms, clustering, dimensionality reduction, and model performance evaluation. By applying these techniques, the study seeks to identify patterns and relationships that contribute to stress management among students.

Through this project, predictive analytics is presented as a powerful decision-support tool that can assist educational institutions, counselors, and policymakers in understanding student needs and designing effective well-being strategies. The work emphasizes the importance of data-driven analysis in addressing real-world problems and showcases how theoretical concepts can be translated into practical solutions using real survey data.

Dataset Description

The dataset used in this project was **self-collected by the author using Google Forms** with the objective of understanding student lifestyle patterns and stress management behavior. The survey was designed to capture multiple dimensions of student life, including academic workload, sleep habits, daily routines, mental well-being, time management, and perceived stress levels. The dataset provides a realistic representation of survey-based data commonly encountered in educational and social analytics.

The dataset consists of **500 observations (rows)**, where each row represents an individual student's response, and **41 attributes (columns)**, where each column corresponds to a distinct survey question. The survey questions were primarily structured using a **Likert scale**, with response options such as *Strongly Disagree, Disagree, Neutral, Agree,* and *Strongly Agree*. This format enables the quantification of subjective opinions and perceptions while maintaining simplicity for respondents.

Since the data was collected through a Google Form, it initially included a **timestamp attribute** indicating the time of submission. This attribute was removed during preprocessing as it did not contribute to the analytical objectives of the study. The dataset does not contain any personally identifiable information such as names, contact details, or student IDs, ensuring respondent anonymity and ethical data usage. Participation in the survey was voluntary, and the data was collected solely for academic research purposes.

The dataset is **categorical in nature**, with most variables representing ordinal data derived from Likert-scale responses. As a result, specialized preprocessing steps such as text normalization, categorical encoding, and missing value handling were required before applying predictive analytics techniques. Minor inconsistencies and missing values were present due to optional questions and variations in user responses, reflecting real-world data collection challenges.

Overall, the dataset serves as a rich and practical foundation for applying **exploratory data analysis, correlation studies, regression, classification, clustering, and dimensionality reduction techniques**. Its structure and size make it well-suited for demonstrating predictive analytics concepts as outlined in the **INT234 course curriculum**, while also providing meaningful insights into student lifestyle and stress-related behaviors.

DATA PREPROCESSING

Data preprocessing is a crucial stage in the predictive analytics pipeline, as the quality of input data directly influences the accuracy and reliability of analytical results. Since the dataset used in this project was collected through a Google Form and consisted primarily of categorical and ordinal responses, several preprocessing steps were required to transform the raw data into a suitable format for analysis and model building.

The first step involved **data cleaning and structure verification**. Column names were standardized by removing leading and trailing spaces to ensure consistency and prevent errors during analysis. The dataset initially contained a timestamp column generated automatically by Google Forms. As this attribute did not provide any analytical or predictive value in the context of student lifestyle and stress analysis, it was permanently removed from the dataset. Additionally, columns that contained only null or empty values were identified and eliminated to reduce noise and improve data quality.

Next, **text normalization** was performed across all remaining attributes. Since survey responses were recorded as textual values, all entries were converted to lowercase, and unnecessary whitespace was removed. This step ensured uniform representation of categorical responses and prevented duplication of categories due to variations in text formatting, such as differences in capitalization or spacing.

Following normalization, **identification of Likert-scale questions** was carried out. The dataset contained multiple survey questions using Likert-scale responses, such as *Strongly Disagree, Disagree, Neutral, Agree,* and *Strongly Agree*. These questions were automatically detected by examining unique response values within each column. Only those columns that followed the Likert-scale pattern were selected for numerical encoding and further analysis.

The identified Likert-scale responses were then **encoded into numerical values** to enable quantitative analysis. A consistent ordinal mapping scheme was applied, where *Strongly Disagree* was mapped to 1, *Disagree* to 2, *Neutral* to 3, *Agree* to 4, and *Strongly Agree* to 5. This encoding preserved the inherent order of responses and allowed the application of correlation analysis, regression models, classification algorithms, and clustering techniques.

During the encoding process, some missing values emerged due to inconsistent or incomplete survey responses. To address this issue, **missing value handling** was performed using a combination of target-aware filtering and mean imputation. Rows with missing values in the selected target variable were removed to maintain modeling integrity, while remaining missing values in predictor variables were replaced with the mean of the respective column. This approach ensured that no missing values were passed to machine learning models, which typically require complete numerical input.

Finally, a **data validation check** was conducted to confirm the absence of null values and verify the suitability of the dataset for predictive modeling. The preprocessed dataset was then ready for

exploratory data analysis, correlation studies, supervised learning, unsupervised learning, and model performance evaluation. These preprocessing steps ensured robustness, minimized bias, and reflected real-world data handling practices commonly followed in data science projects.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split, KFold, cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
```

Data Loading & Cleaning

[2]:

```
df = pd.read_csv("Student Lifestyle & Well-being Survey (Responses) - Form Responses 1 (1).csv")

# Clean column names
df.columns = df.columns.str.strip()

# Remove Timestamp column completely
df = df.loc[:, ~df.columns.str.contains("timestamp", case=False)]

# Remove fully empty columns
df = df.dropna(axis=1, how="all")

# Normalize all values (survey-safe)
for col in df.columns:
    df[col] = df[col].astype(str).str.strip().str.lower()

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 502 entries, 0 to 501
Data columns (total 42 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   What is your current age?
502 non-null    object
 1   What is your gender identity?
502 non-null    object
 2   Which year of study are you currently in?
502 non-null    object
 3   Do you live on-campus (e.g., dormitory/hostel)?
502 non-null    object
 4   On a typical weekday, how many hours do you spend on academic study (classes, homework, studying)?
502 non-null    object
 5   I feel overwhelmed by my academic workload.
```

EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis (EDA) is a fundamental step in predictive analytics that focuses on understanding the structure, distribution, and underlying patterns present in the dataset before applying advanced modeling techniques. The primary objective of EDA in this project was to gain insights into student lifestyle behaviors, stress perception, and response trends, as well as to identify anomalies, dominant patterns, and potential relationships among variables.

Since the dataset consists mainly of categorical and ordinal variables derived from Likert-scale responses, bar charts were chosen as the primary visualization technique. Bar charts are effective for representing frequency distributions of categorical data and provide a clear visual understanding of how students responded to different survey questions. For each relevant question, response frequencies were plotted to analyze trends related to academic workload, sleep habits, time management, mental well-being, and stress management.

The EDA revealed that a significant proportion of students reported experiencing moderate to high levels of academic pressure. Responses related to sleep patterns indicated that many students do not maintain a consistent or adequate sleep schedule, which is a known contributor to increased stress levels. Questions associated with time management and daily routine balance showed varied responses, suggesting differences in students' ability to manage academic and personal responsibilities effectively.

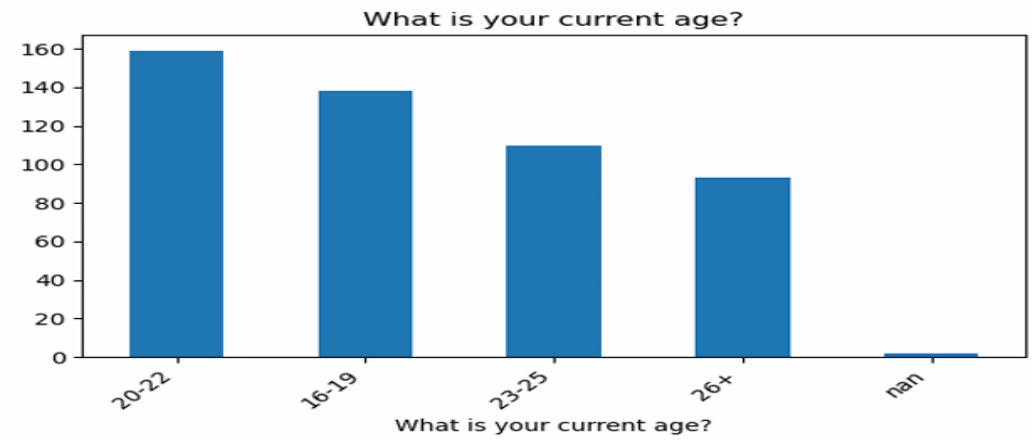
Visual analysis of stress-related questions highlighted that while some students felt confident in managing their stress, a notable percentage expressed difficulty in handling stress effectively. This variation justified the selection of stress management as a key analytical focus in the study. The distribution of responses across multiple lifestyle-related questions indicated that stress is not influenced by a single factor but is instead the result of a combination of academic, behavioral, and lifestyle elements.

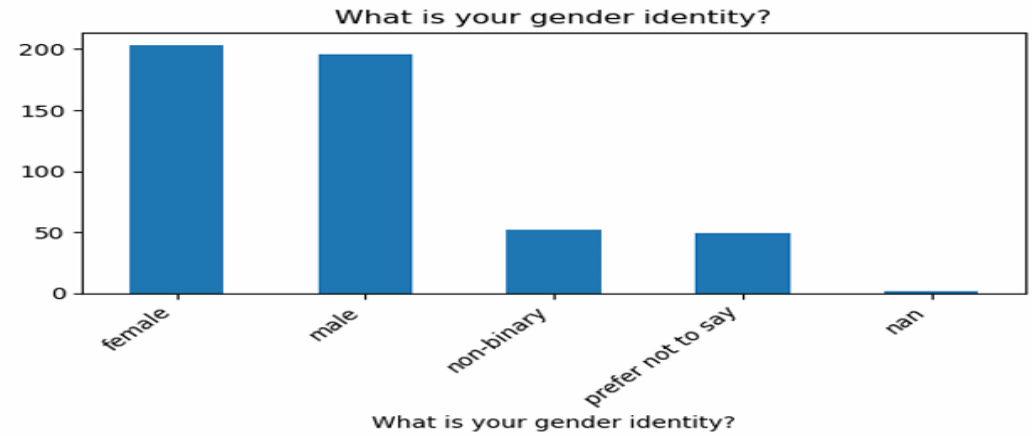
EDA also helped identify minor inconsistencies and missing responses in certain questions, which were subsequently addressed during the data preprocessing stage. No extreme outliers were observed due to the ordinal nature of the Likert-scale data. However, differences in response concentration across categories provided meaningful insights into dominant behavioral patterns within the student population.

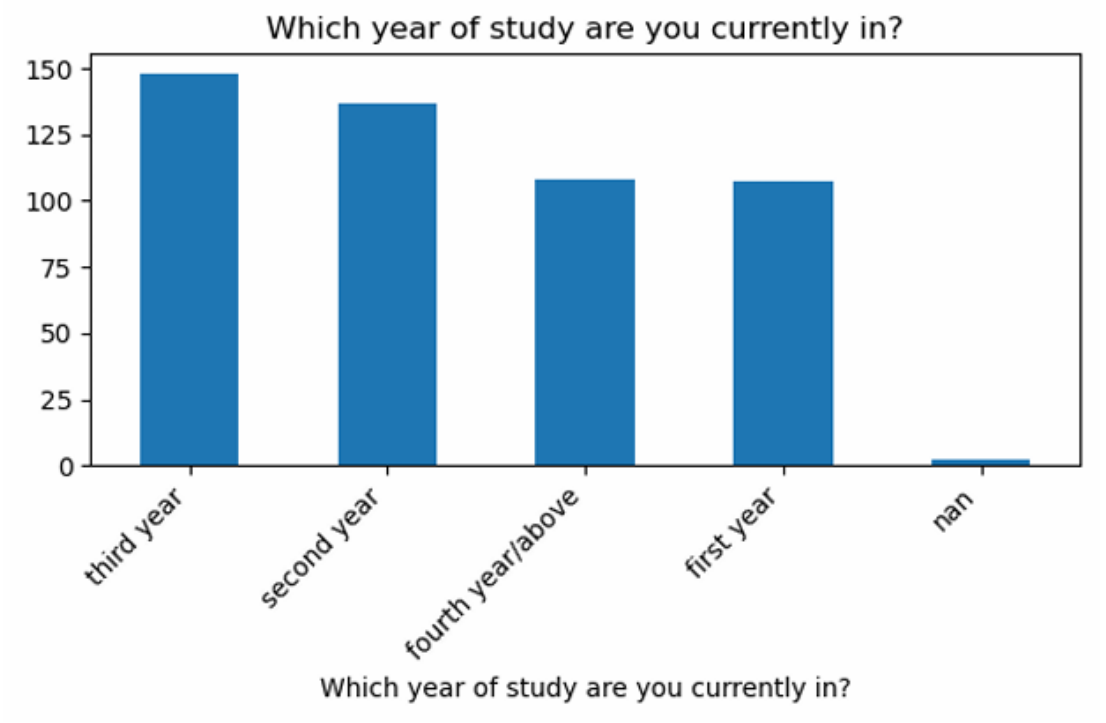
Overall, Exploratory Data Analysis played a crucial role in shaping the direction of further analysis. It provided a strong foundation for correlation analysis, regression modeling, classification, and clustering by offering a clear understanding of response distributions and key trends in student lifestyle and stress-related behavior. The insights derived from EDA reinforced the importance of adopting a data-driven approach to assess student well-being and guided the selection of appropriate predictive analytics techniques used in subsequent stages of the project.

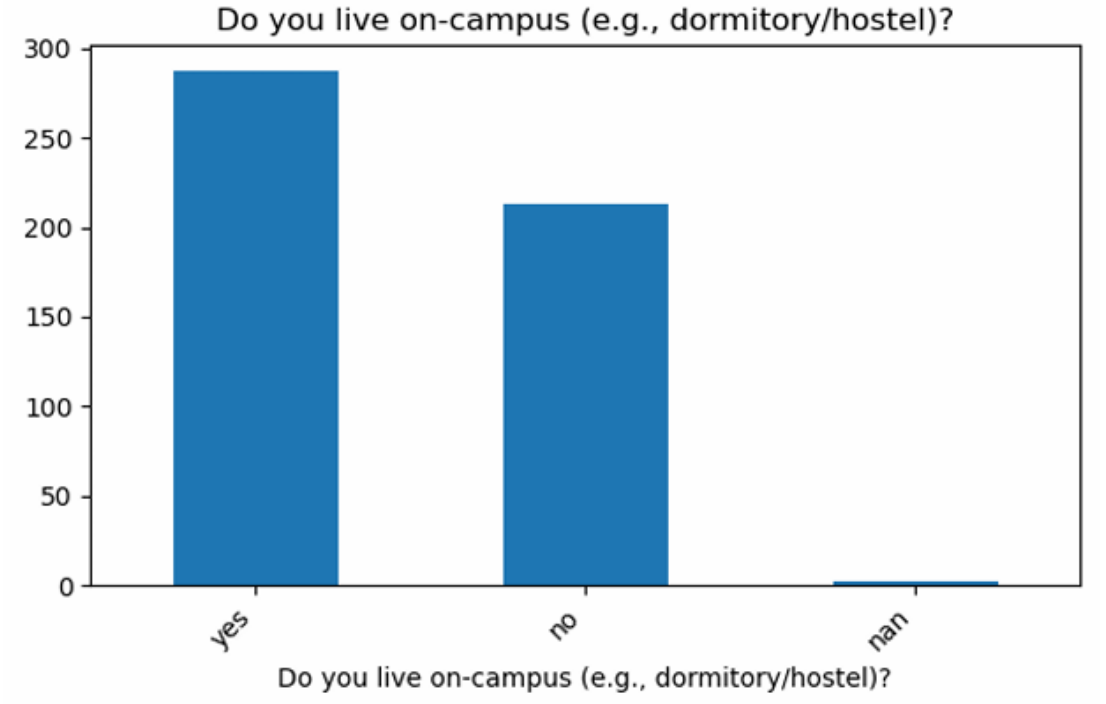
Exploratory Data Analysis

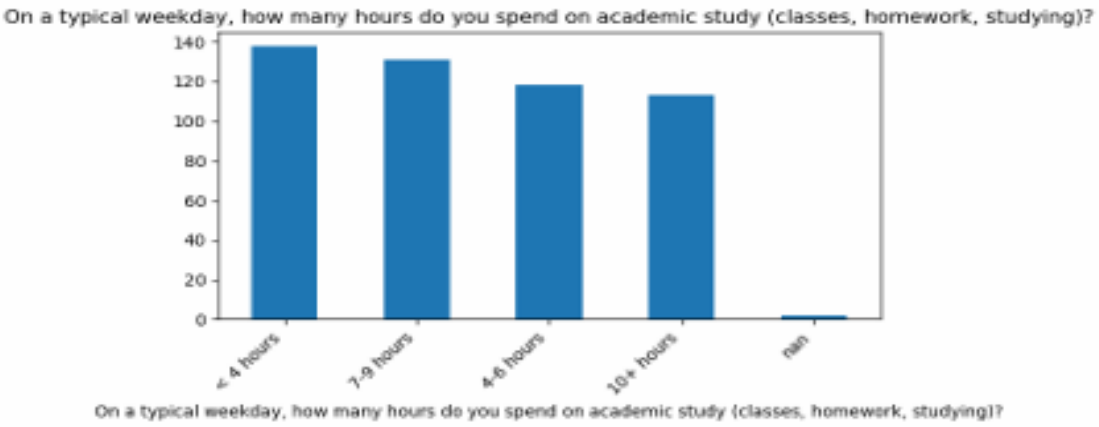
```
def plot_bar(col):
    counts = df[col].value_counts()
    if counts.empty:
        return
    plt.figure(figsize=(6,4))
    counts.plot(kind="bar")
    plt.title(col)
    plt.xticks(rotation=45, ha="right")
    plt.tight_layout()
    plt.show()

for col in df.columns:
    plot_bar(col)
```

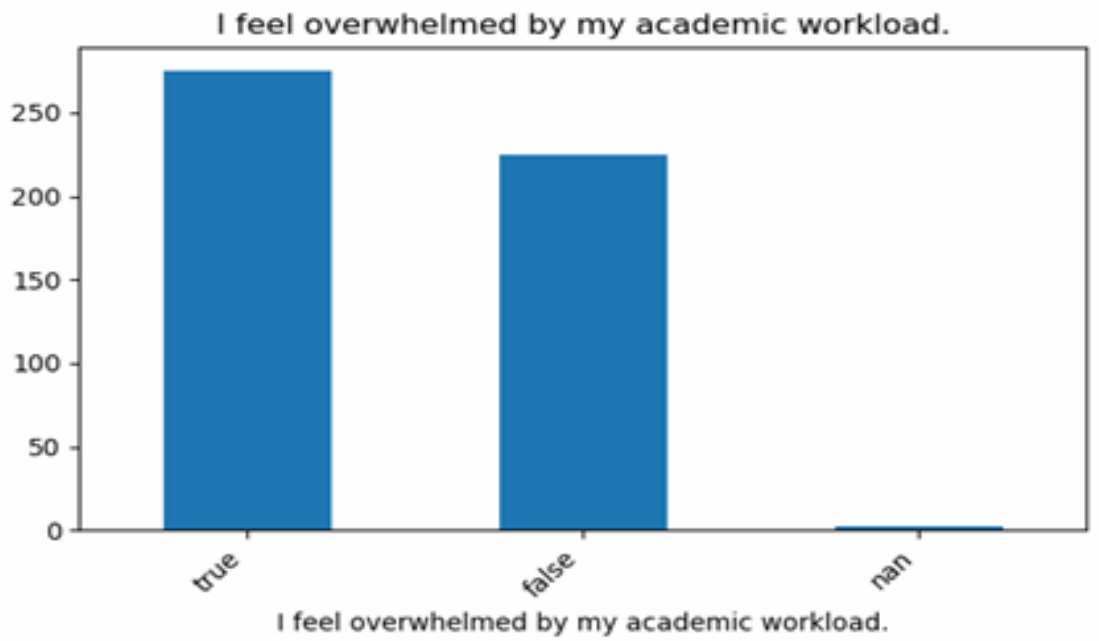


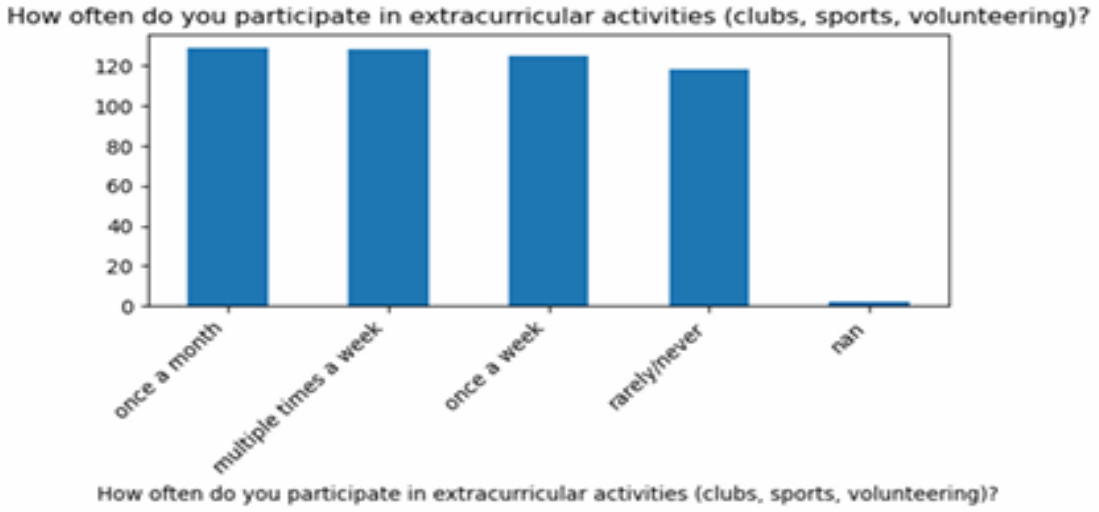


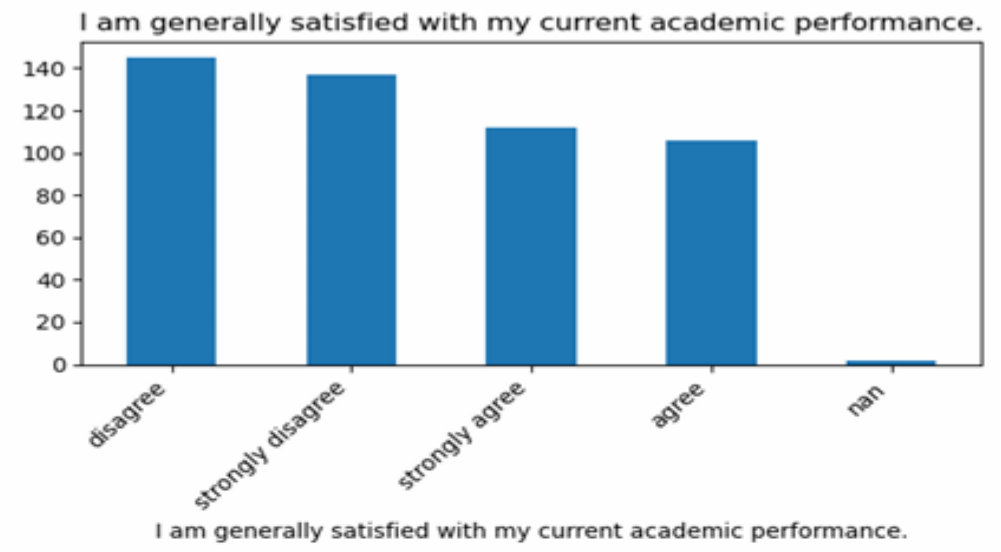

Which year of study are you currently in?
150
125
100
75
50
25
0
third year
second year
fourth year/above
first year
nan
Which year of study are you currently in?

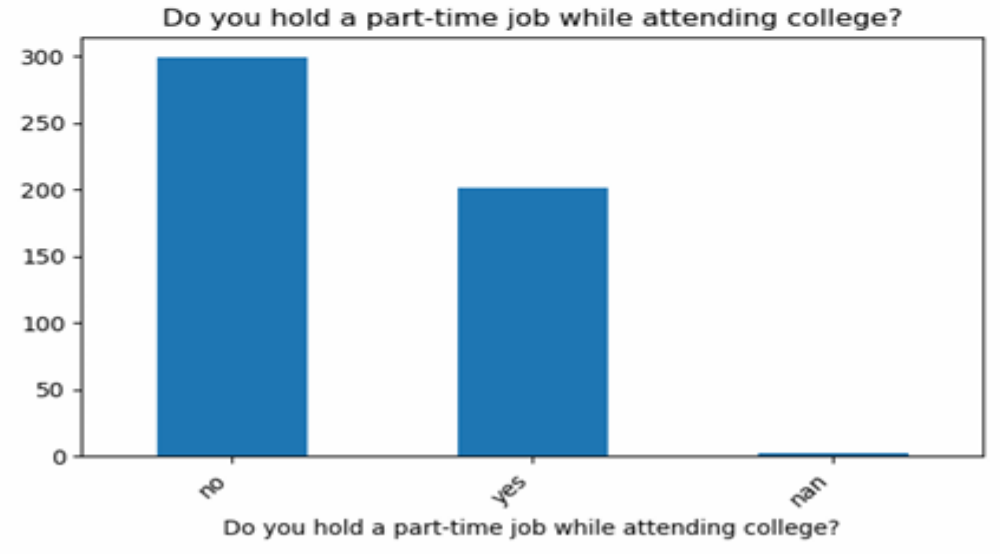

Do you live on-campus (e.g., dormitory/hostel)?
300
250
200
150
100
50
0
yes
no
nan
Do you live on-campus (e.g., dormitory/hostel)?

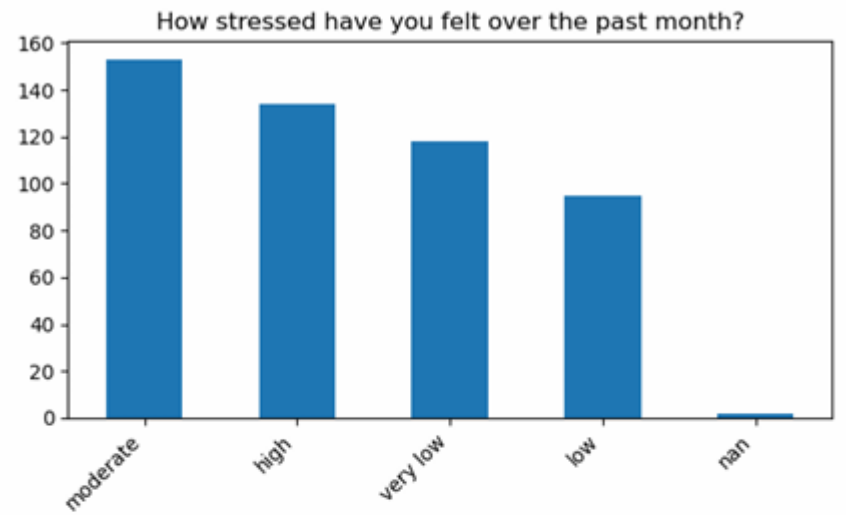

On a typical weekday, how many hours do you spend on academic study (classes, homework, studying)?
140
120
100
80
60
40
20
0
< 4 hours
7-9 hours
4-6 hours
10+ hours
nan
On a typical weekday, how many hours do you spend on academic study (classes, homework, studying)?

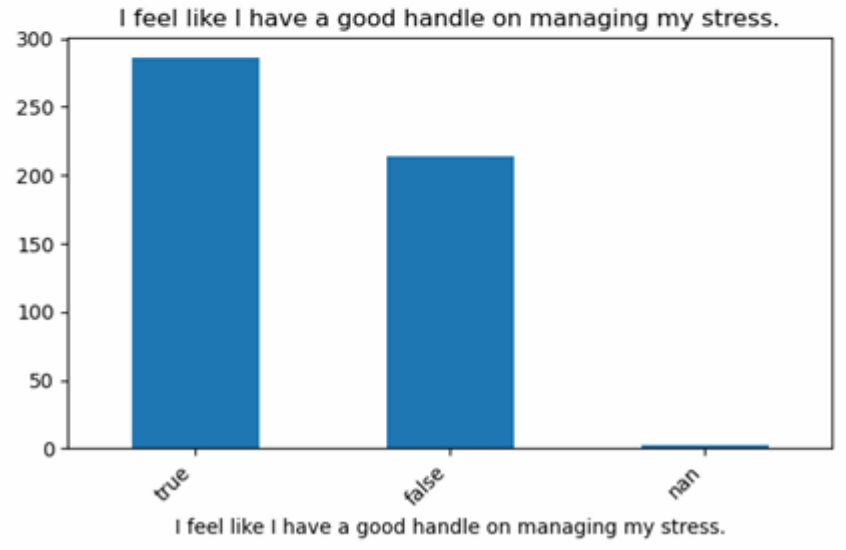

I feel overwhelmed by my academic workload.
250
200
150
100
50
0
true
false
nan
I feel overwhelmed by my academic workload.

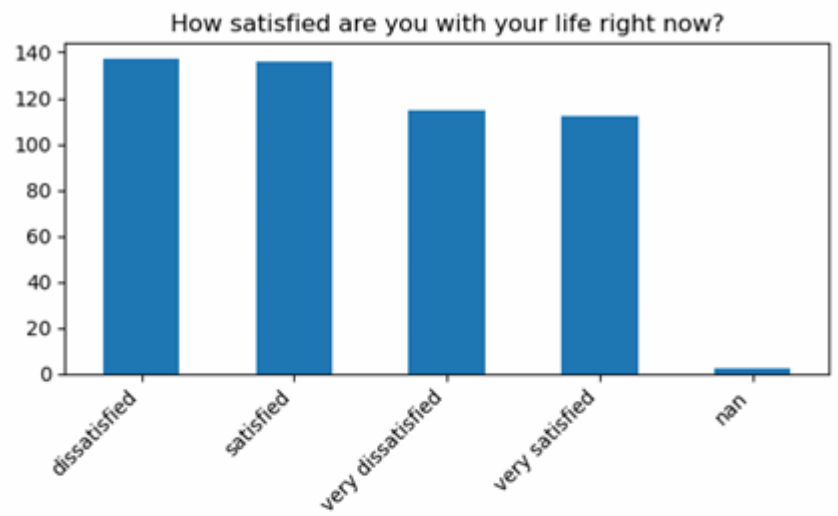

How often do you participate in extracurricular activities (clubs, sports, volunteering)?
120
100
80
60
40
20
0
once a month
multiple times a week
once a week
rarely/never
nan
How often do you participate in extracurricular activities (clubs, sports, volunteering)?


I am generally satisfied with my current academic performance.
140
120
100
80
60
40
20
0
disagree
strongly disagree
strongly agree
agree
nan
I am generally satisfied with my current academic performance.

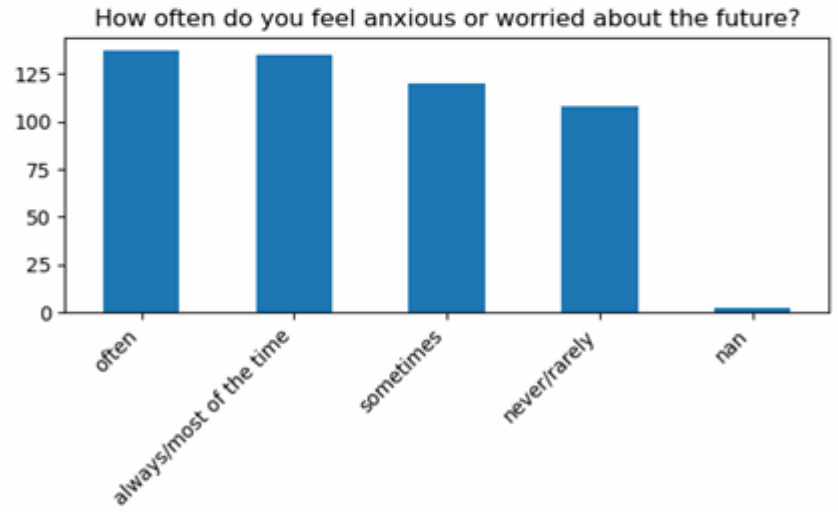

Do you hold a part-time job while attending college?
300
250
200
150
100
50
0
no
yes
nan
Do you hold a part-time job while attending college?


How stressed have you felt over the past month?
160
140
120
100
80
60
40
20
0
moderate
high
very low
low
nan
How stressed have you felt over the past month?

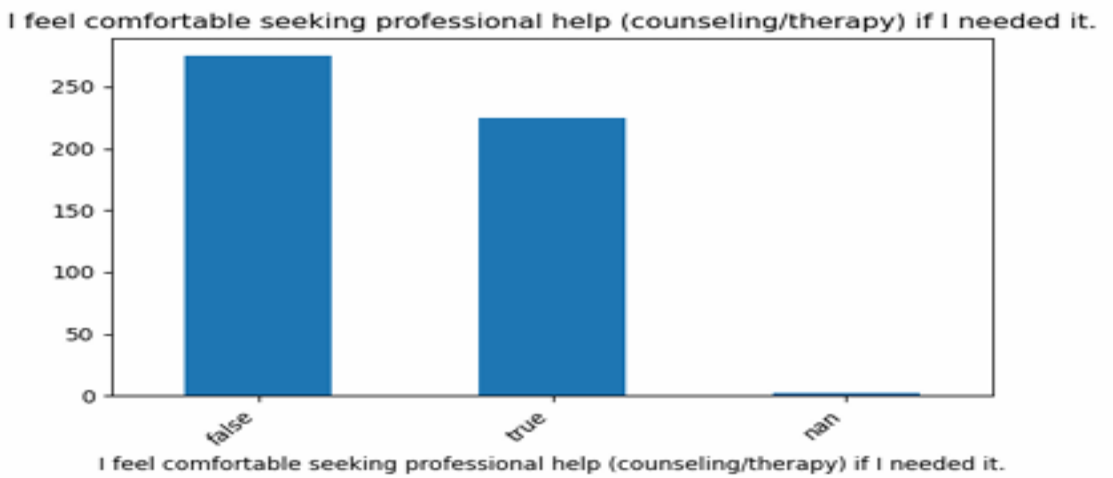

I feel like I have a good handle on managing my stress.
300
250
200
150
100
50
0
true
false
nan
I feel like I have a good handle on managing my stress.

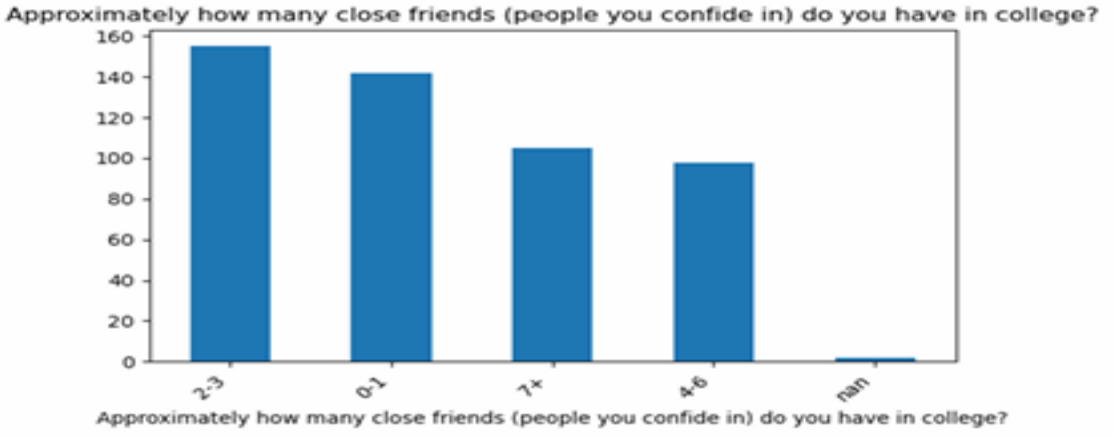


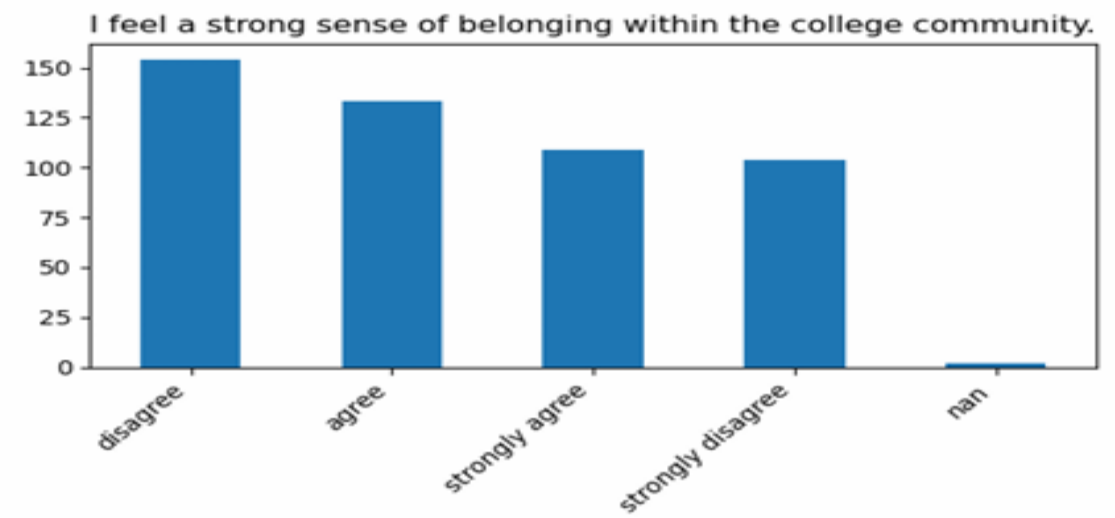
How satisfied are you with your life right now?



How often do you feel anxious or worried about the future?

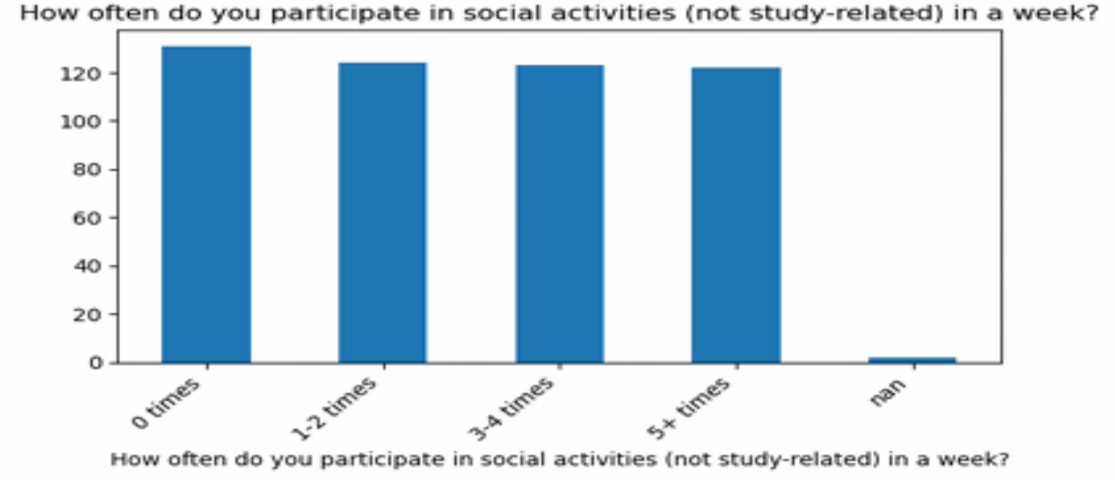

I feel comfortable seeking professional help (counseling/therapy) if I needed it.
250
200
150
100
50
0
false
true
nan
I feel comfortable seeking professional help (counseling/therapy) if I needed it.


Approximately how many close friends (people you confide in) do you have in college?
160
140
120
100
80
60
40
20
0
2-3
0-1
7+
4-6
nan
Approximately how many close friends (people you confide in) do you have in college?

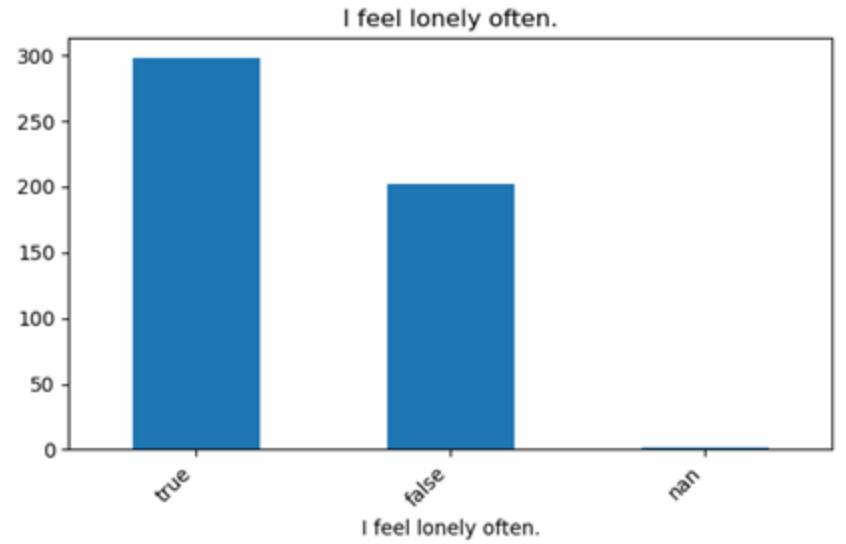


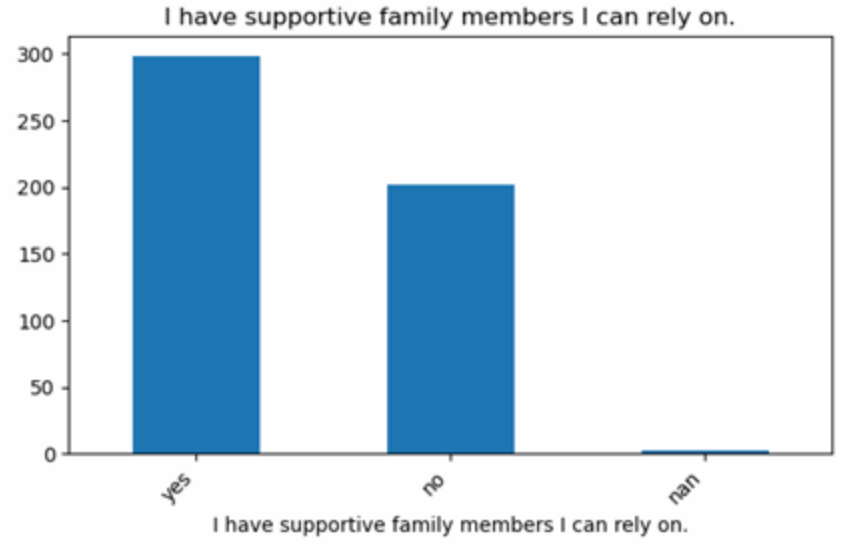
I feel a strong sense of belonging within the college community.

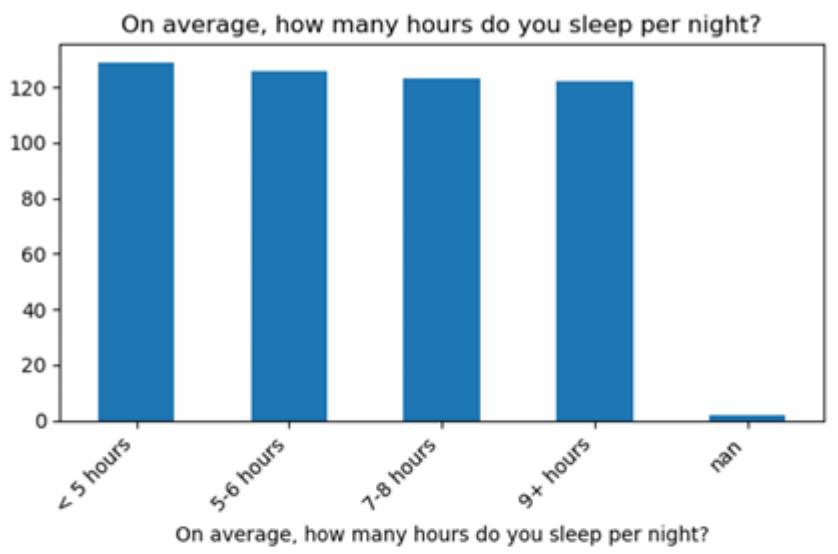
How often do you participate in social activities (not study-related) in a week?

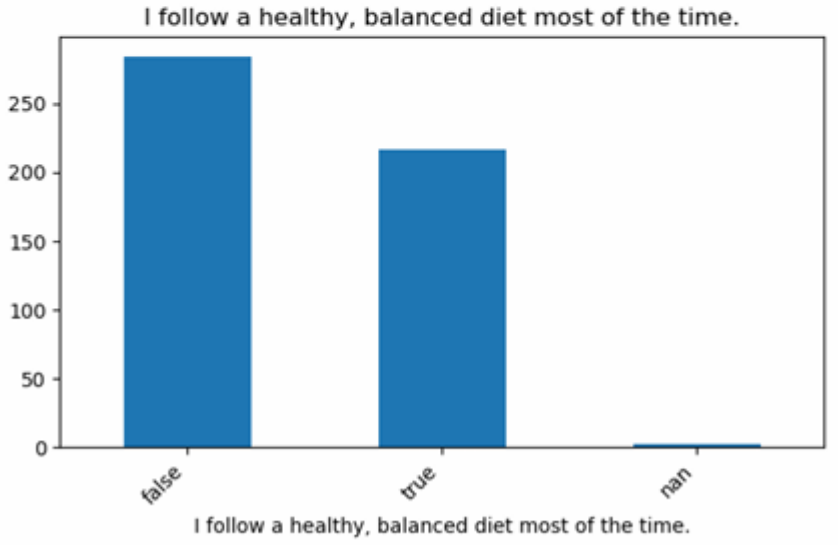


How often do you participate in social activities (not study-related) in a week?


I feel lonely often.
300
250
200
150
100
50
0
true
false
nan
I feel lonely often.

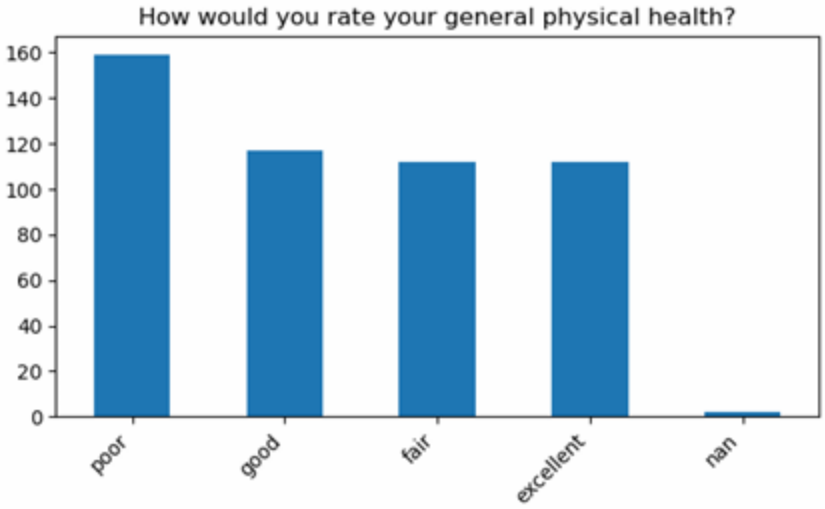

I have supportive family members I can rely on.
300
250
200
150
100
50
0
yes
no
nan
I have supportive family members I can rely on.


On average, how many hours do you sleep per night?
120
100
80
60
40
20
0
< 5 hours
5-6 hours
7-8 hours
9+ hours
nan
On average, how many hours do you sleep per night?


I follow a healthy, balanced diet most of the time.
250
200
150
100
50
0
false
true
nan
I follow a healthy, balanced diet most of the time.

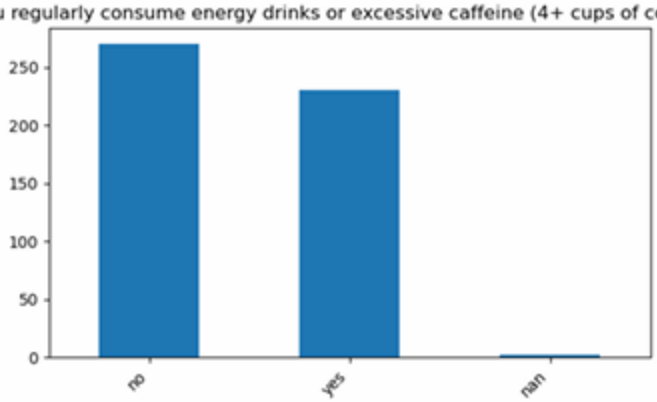
How often do you engage in vigorous physical exercise (e.g., running, gym, sports) per week?

How often do you engage in vigorous physical exercise (e.g., running, gym, sports) per week?



How would you rate your general physical health?

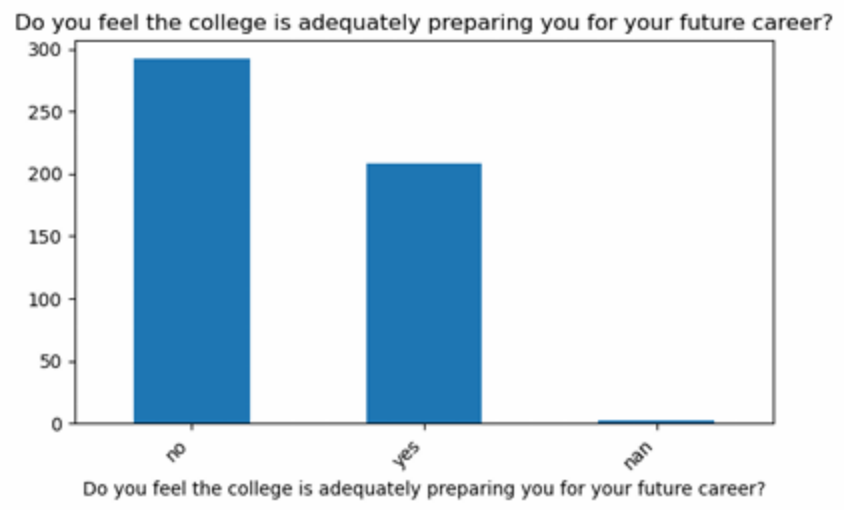
Do you regularly consume energy drinks or excessive caffeine (4+ cups of coffee/day)?

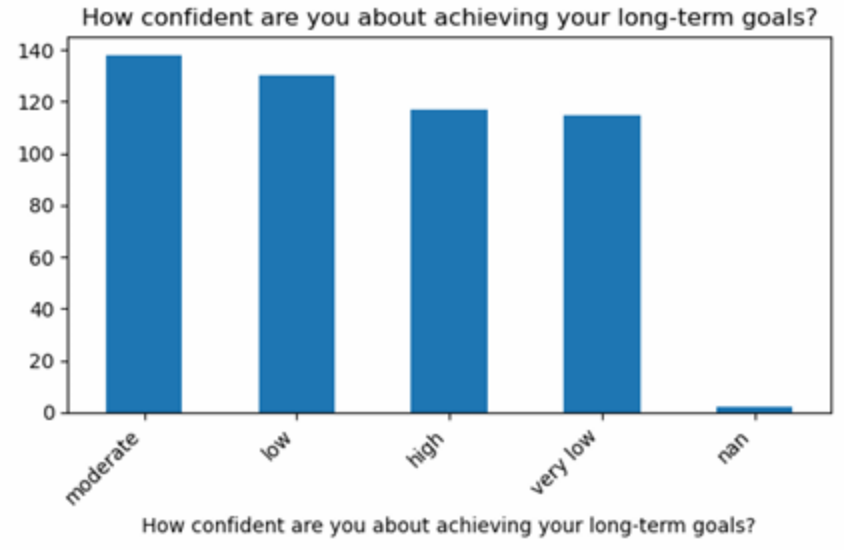


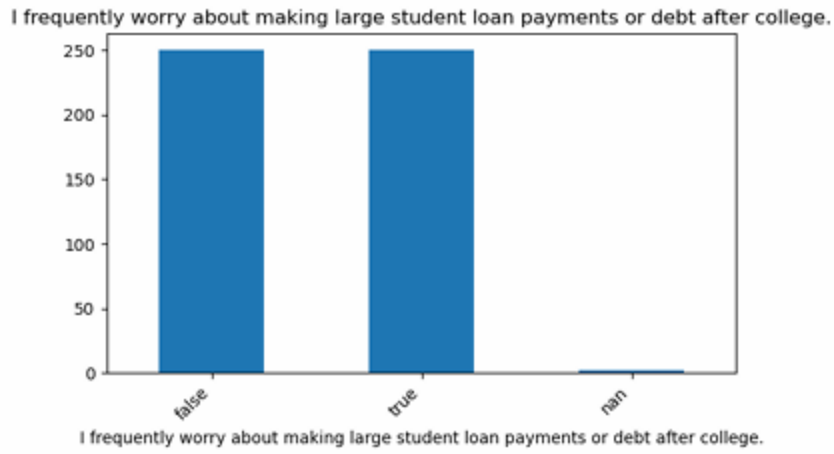
Do you regularly consume energy drinks or excessive caffeine (4+ cups of coffee/day)?

I am worried about my current financial situation.

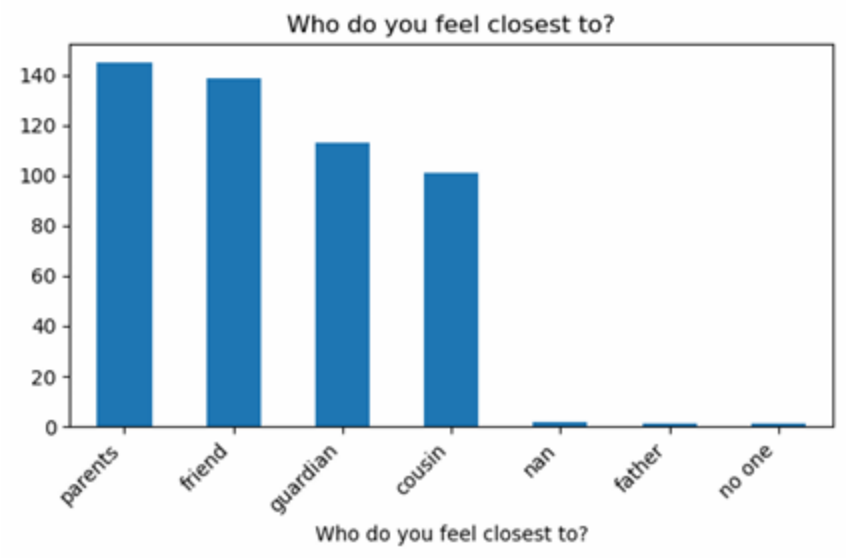
I feel prepared for my career or life after graduation.

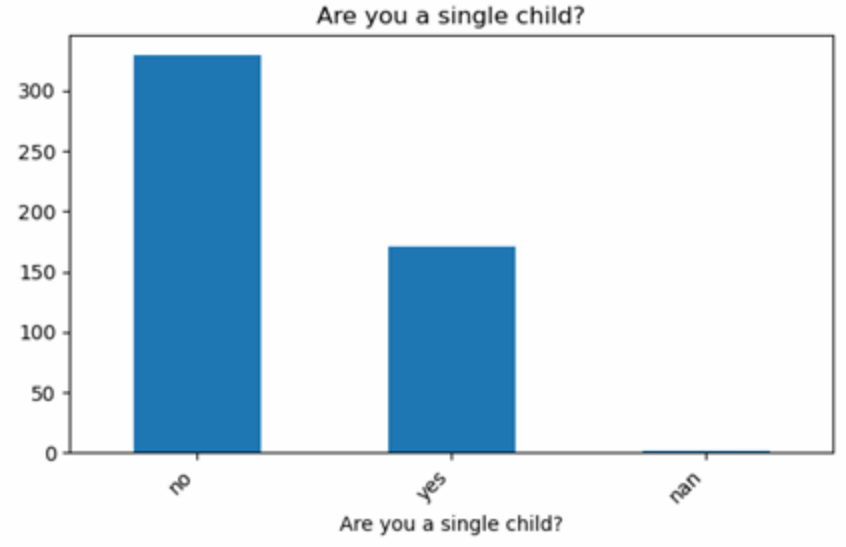

Do you feel the college is adequately preparing you for your future career?
300
250
200
150
100
50
0
no
yes
nan
Do you feel the college is adequately preparing you for your future career?

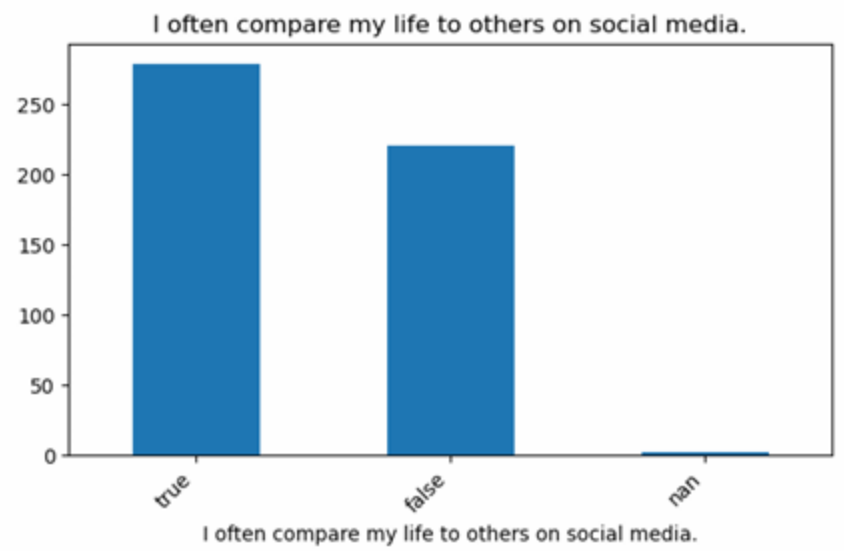

How confident are you about achieving your long-term goals?
140
120
100
80
60
40
20
0
moderate
low
high
very low
nan
How confident are you about achieving your long-term goals?

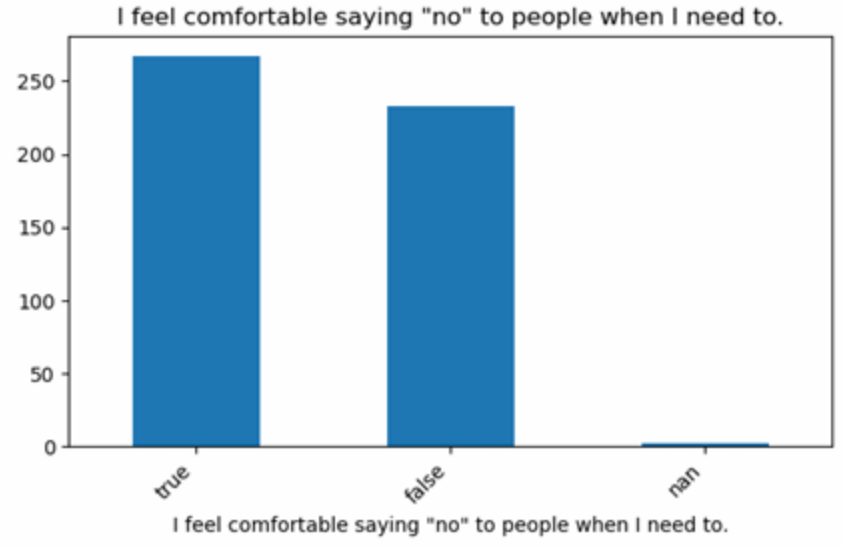

I frequently worry about making large student loan payments or debt after college.
250
200
150
100
50
0
false
true
nan
I frequently worry about making large student loan payments or debt after college.


Do you regularly track your personal spending/budget?
300
250
200
150
100
50
0
no
yes
nan
Do you regularly track your personal spending/budget?

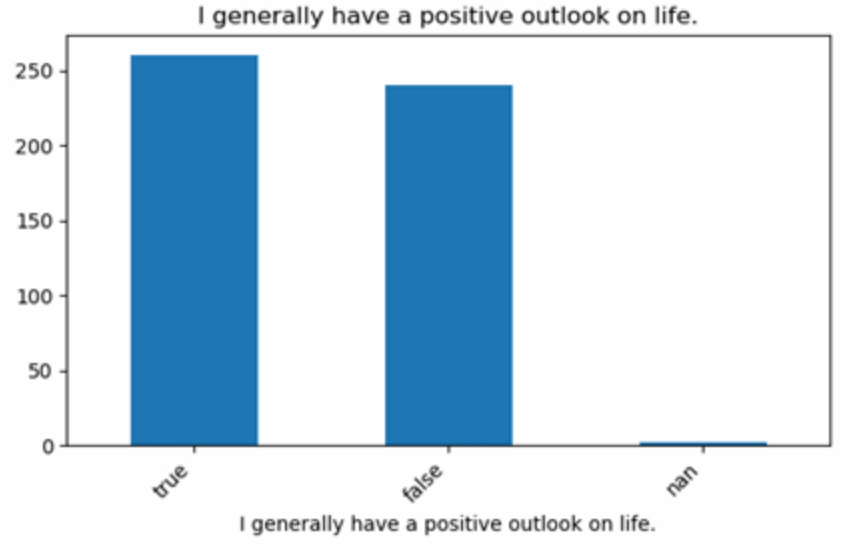

Who do you feel closest to?
140
120
100
80
60
40
20
0
parents
friend
guardian
cousin
nan
father
no one
Who do you feel closest to?

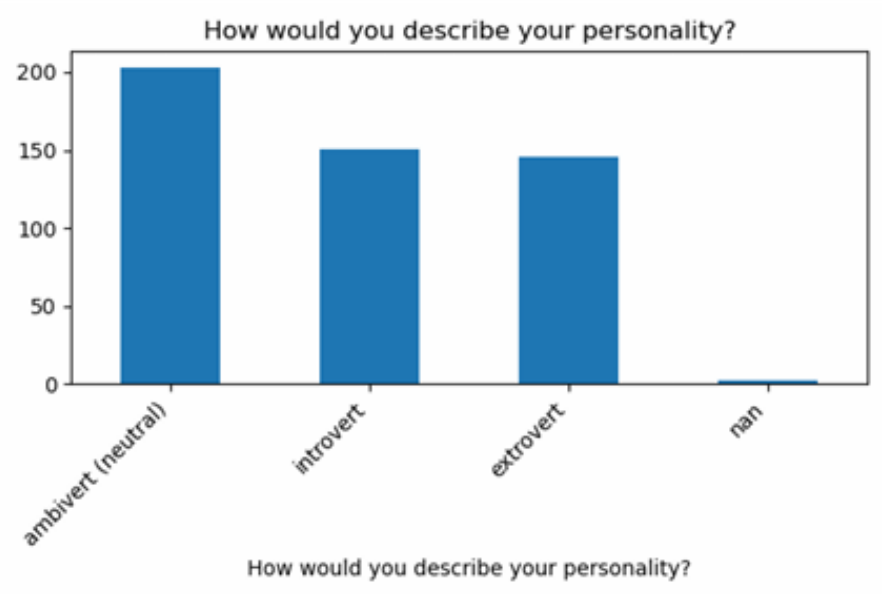

Are you a single child?
300
250
200
150
100
50
0
no
yes
nan
Are you a single child?

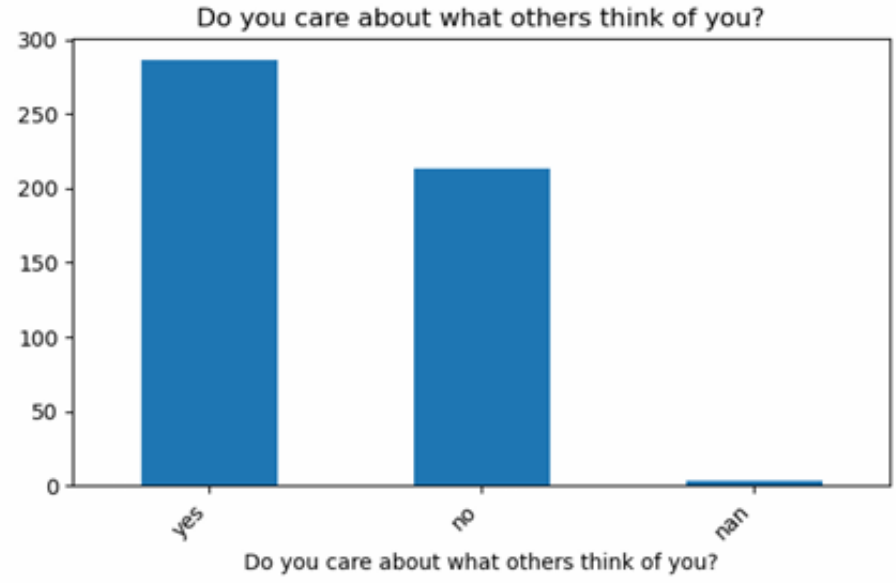

I often compare my life to others on social media.
250
200
150
100
50
0
true
false
nan
I often compare my life to others on social media.


I feel comfortable saying "no" to people when I need to.
250
200
150
100
50
0
true
false
nan
I feel comfortable saying "no" to people when I need to.


Do you have a clear set of goals for the next 5 years?
250
200
150
100
50
0
no
yes
nan
Do you have a clear set of goals for the next 5 years?


I generally have a positive outlook on life.
250
200
150
100
50
0
true
false
nan
I generally have a positive outlook on life.


How would you describe your personality?
200
150
100
50
0
ambivert (neutral)
introvert
extrovert
nan
How would you describe your personality?


Do you care about what others think of you?
300
250
200
150
100
50
0
yes
no
nan
Do you care about what others think of you?

CORRELATION ANALYSIS

Correlation analysis is an important statistical technique used to measure the strength and direction of relationships between variables. In this project, correlation analysis was performed to understand how different aspects of student lifestyle are related to stress management and overall well-being. Since the dataset primarily consisted of Likert-scale responses, ordinal variables were first encoded into numerical values to enable quantitative correlation analysis.

After encoding, a **correlation matrix** was computed using the Pearson correlation coefficient. This method was selected due to its widespread use in measuring linear relationships between numerical variables. The correlation values range from −1 to +1, where positive values indicate a direct relationship, negative values indicate an inverse relationship, and values close to zero suggest little or no linear association. A **correlation heatmap** was generated to visually represent the relationships among the selected lifestyle and stress-related variables.

The correlation analysis revealed several meaningful patterns within the dataset. Variables related to effective time management, adequate sleep, and balanced daily routines showed a **positive correlation** with students' perceived ability to manage stress. This suggests that students who maintain healthier lifestyle habits are more likely to report better stress management. Conversely, variables associated with academic overload and irregular routines demonstrated a **negative correlation** with stress management indicators, highlighting the impact of excessive workload on student well-being.

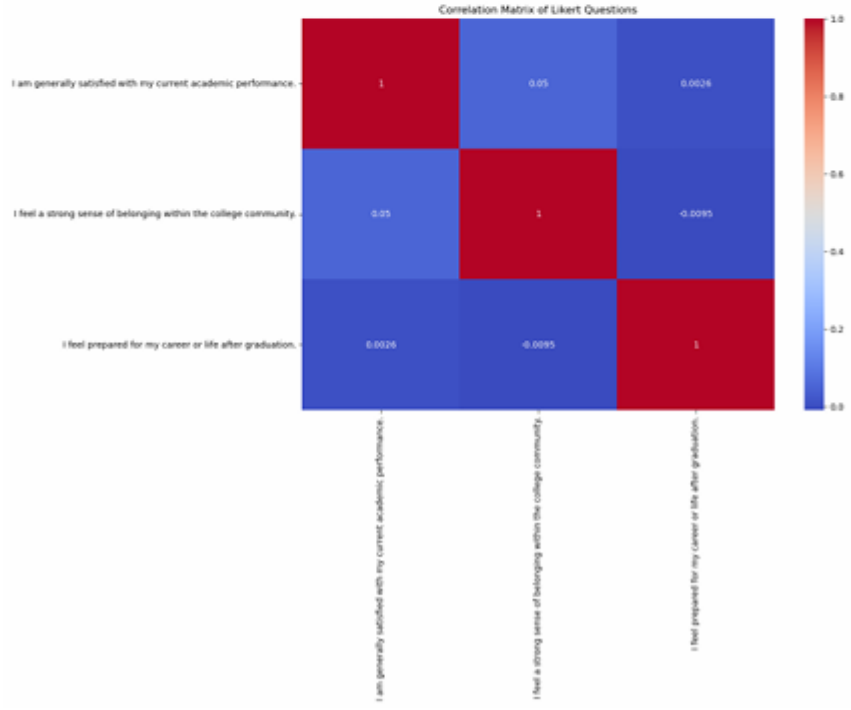
While some correlations were moderate in strength, the results indicate that stress management is influenced by multiple interrelated factors rather than a single dominant variable. The analysis also confirmed that no extremely high correlations existed among predictor variables, reducing the risk of multicollinearity in subsequent regression models. This validation step was important to ensure the reliability and interpretability of supervised learning techniques applied later in the project.

Overall, correlation analysis provided valuable insights into the relationships between student lifestyle behaviors and stress levels. The findings supported the hypothesis that lifestyle factors collectively influence stress management and justified the use of predictive modeling techniques such as regression and classification. By identifying significant associations, this analysis helped guide feature selection and strengthened the analytical foundation of the study.

Correlation Analysis

In [6]:
```
if numeric_df.shape[1] >= 2:
    corr = numeric_df.corr()
    plt.figure(figsize=(12,8))
    sns.heatmap(corr, annot=True, cmap="coolwarm")
    plt.title("Correlation Matrix of Likert Questions")
    plt.show()
else:
    print("Correlation skipped: insufficient numeric variables.")
```


25, 8:41 PM j



REGRESSION ANALYSIS

Regression analysis is a supervised learning technique used to model the relationship between a dependent variable and one or more independent variables. In this project, regression analysis was employed to examine how various lifestyle and behavioral factors influence students' ability to manage stress. The objective was to quantify the impact of different predictors on stress-related outcomes and to assess the predictive capability of the regression model.

To perform regression analysis, a **stress-related Likert-scale question** was selected as the target variable. This variable represents students' self-perceived ability to handle stress. The remaining encoded lifestyle-related variables, such as sleep habits, time management, academic workload, and daily routines, were treated as independent variables. Prior to model training, rows containing missing values in the target variable were removed, and remaining missing values in predictor variables were handled using mean imputation to ensure model compatibility.

A **Multiple Linear Regression model** was implemented to capture the combined effect of multiple predictors on stress management. The dataset was divided into training and testing sets using a standard train-test split approach to evaluate model performance on unseen data. The regression model was trained on the training set and tested on the testing set to assess its generalization ability.

Model performance was evaluated using standard regression metrics, including **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, **Root Mean Squared Error (RMSE)**, and **R-squared (R^2) score**. MAE and RMSE provided insights into the average magnitude of prediction errors, while the R^2 score indicated the proportion of variance in the target variable explained by the model. The obtained results demonstrated that lifestyle-related variables collectively contribute to predicting stress management, although individual factors may vary in their influence.

The regression analysis highlighted that students with better time management skills, healthier sleep patterns, and balanced daily routines tend to exhibit higher stress management scores. Conversely, increased academic pressure and irregular habits were associated with lower predicted stress management levels. These findings align with the trends observed during exploratory data analysis and correlation analysis, reinforcing the validity of the results.

Overall, regression analysis served as an effective tool for understanding the relationship between student lifestyle factors and stress management. It provided quantitative evidence supporting the influence of lifestyle behaviors on stress levels and established a foundation for further supervised learning techniques, such as classification, explored in subsequent sections of the project.

Regression Modeling (NaN-Safe)

In [8]:
```
# Prepare X and y safely
X = numeric_df.drop(columns=[stress_col])
y = numeric_df[stress_col]

# Drop rows with missing target
mask = y.notna()
X = X.loc[mask]
y = y.loc[mask]
```



```
# Final imputation for X
X = X.fillna(X.mean())

print("Remaining NaN in X:", X.isnull().sum().sum())
print("Remaining NaN in y:", y.isnull().sum())

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42
)

lr = LinearRegression()
lr.fit(X_train, y_train)

y_pred = lr.predict(X_test)

print("MAE:", mean_absolute_error(y_test, y_pred))
print("MSE:", mean_squared_error(y_test, y_pred))
print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred)))
print("R²:", r2_score(y_test, y_pred))
```

```
Remaining NaN in X: 0
Remaining NaN in y: 0
MAE: 1.4919430353300251
MSE: 2.5172269826325464
RMSE: 1.5865771278549765
R²: -0.005883531345211024
```

CLASSIFICATION ANALYSIS

Classification analysis is a supervised learning approach used to categorize data into predefined classes. In this project, classification was applied to distinguish students based on their ability to manage stress effectively. The objective of this analysis was to predict whether a student falls into a category of **effective stress management** or **ineffective stress management** based on lifestyle-related attributes.

To perform classification, the stress-related target variable was transformed into a **binary class**. Responses indicating agreement or strong agreement with effective stress handling were categorized as one class, while neutral or disagreement responses were categorized as the other. This transformation enabled the application of classification algorithms while maintaining interpretability.

A **Logistic Regression classifier** was implemented due to its simplicity, interpretability, and suitability for binary classification problems. The dataset was divided into training and testing subsets to evaluate the model's predictive performance on unseen data. Prior to model training, necessary preprocessing steps ensured that no missing values were present in the input features.

The performance of the classification model was evaluated using multiple metrics, including **Accuracy, Precision, Recall, F1-score, and Confusion Matrix**. Accuracy measured the overall correctness of predictions, while precision and recall provided insights into the model's ability to correctly identify students with effective stress management. The confusion matrix visually summarized correct and incorrect predictions across classes.

The classification results demonstrated that lifestyle-related variables contain sufficient information to reasonably predict stress management categories. This analysis supports the use of predictive analytics for early identification of students who may require additional academic or mental health support.

Classification (Binary Stress Handling)

In [9]:
```
y_class = (y >= 4).astype(int)

X_train, X_test, y_train, y_test = train_test_split(
    X, y_class, test_size=0.3, random_state=42
)

clf = LogisticRegression(max_iter=1000)
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)

print("Accuracy:", accuracy_score(y_test, y_pred))
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.5894039735099338
[[89  0]
 [62  0]]
              precision    recall  f1-score   support

           0       0.59      1.00      0.74        89
           1       0.00      0.00      0.00        62

    accuracy                           0.59       151
   macro avg       0.29      0.50      0.37       151
weighted avg       0.35      0.59      0.44       151
```

DIMENSIONALITY REDUCTION USING PCA

Dimensionality reduction is an important technique used to simplify high-dimensional data while preserving essential information. In this project, **Principal Component Analysis (PCA)** was applied to reduce the number of features and visualize student clusters more effectively.

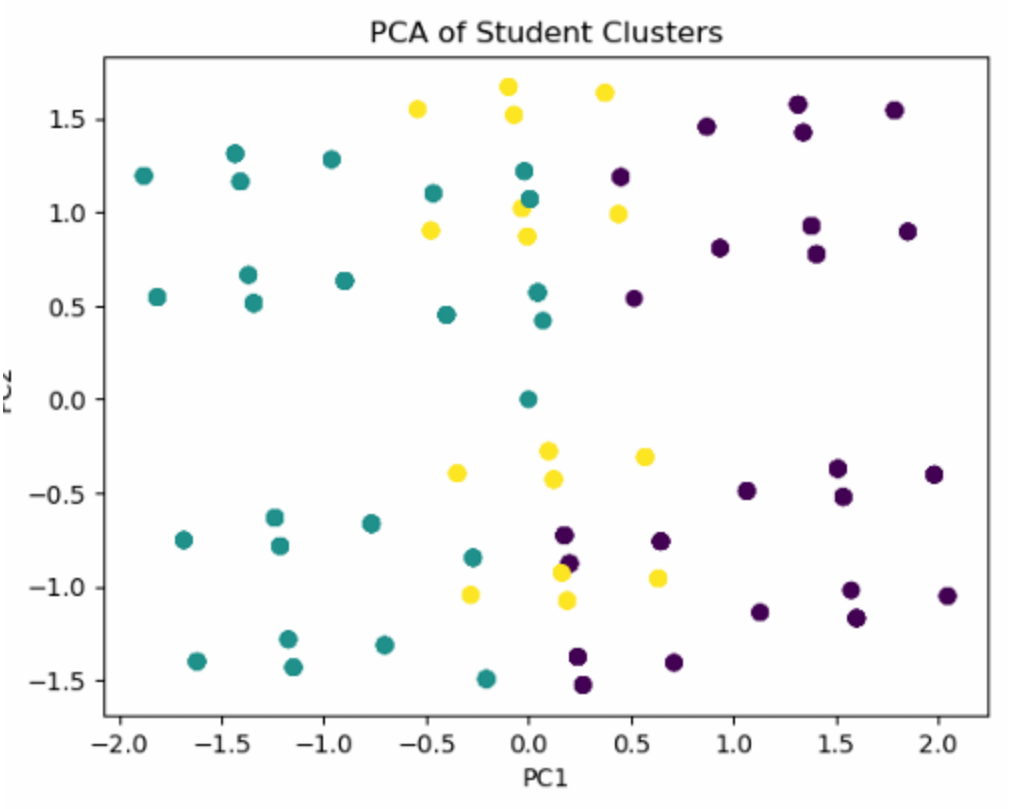
PCA works by transforming the original variables into a smaller set of uncorrelated components known as principal components. These components capture the maximum variance present in the data. After standardizing the dataset, PCA was applied to project the data into two principal components, enabling graphical visualization.

The PCA visualization provided a clear representation of student clusters identified through K-Means clustering. It highlighted the separation between different student groups based on lifestyle and stress-related features. This visualization helped validate the clustering results and offered an intuitive understanding of complex, multi-dimensional survey data.

Thus, PCA proved to be a valuable tool for dimensionality reduction, data visualization, and pattern interpretation within the context of this project.

Clustering & PCA

[10]:
```
scaler = StandardScaler()
scaled = scaler.fit_transform(numeric_df)

kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(scaled)

pca = PCA(n_components=2)
pca_data = pca.fit_transform(scaled)

plt.scatter(pca_data[:,0], pca_data[:,1], c=clusters)
plt.xlabel("PC1")
plt.ylabel("PC2")
plt.title("PCA of Student Clusters")
plt.show()
```



MODEL EVALUATION (K-FOLD CROSS VALIDATION)

Model evaluation is essential to assess the reliability and generalizability of predictive models. In this project, **K-Fold Cross Validation** was used to evaluate the performance of the regression model. This technique divides the dataset into multiple subsets, or folds, and trains the model repeatedly using different combinations of training and testing data.

K-Fold Cross Validation helps reduce the impact of random data splits and provides a more robust estimate of model performance. The regression model was evaluated across multiple folds, and performance metrics were averaged to assess consistency. This approach also helps address the **bias–variance trade-off**, ensuring that the model neither overfits nor underfits the data.

The cross-validation results indicated stable performance across different folds, suggesting that the model generalizes well to unseen data. This evaluation technique strengthened the credibility of the predictive analytics pipeline used in the project.

Model Performance: K-Fold Cross Validation

[11]:
```
kf = KFold(n_splits=5, shuffle=True, random_state=42)
scores = cross_val_score(lr, X, y, cv=kf, scoring="r2")

print("Cross-validation R² scores:", scores)
print("Mean R²:", scores.mean())
```

```
Cross-validation R² scores: [-0.04027186 -0.04596709 -0.01047646  0.00426044 -0.0
7402009]
Mean R²: -0.033295009795312636
```

CONCLUSION

This project successfully demonstrated the application of predictive analytics techniques to analyze student lifestyle and stress management using a real-world, self-collected survey dataset. Starting from data preprocessing and exploratory analysis, the study applied supervised and unsupervised learning methods to uncover meaningful insights into student behavior.

Exploratory Data Analysis and correlation studies revealed key trends and relationships among lifestyle factors and stress levels. Regression analysis quantified the influence of lifestyle attributes on stress management, while classification techniques enabled prediction of stress management categories. Clustering and PCA further enhanced understanding by identifying distinct student groups and visualizing complex data patterns. Model evaluation using K-Fold Cross Validation ensured robustness and reliability of results.

Overall, the project highlights the effectiveness of predictive analytics in addressing real-life problems in the educational domain. The findings emphasize the importance of balanced lifestyle habits for effective stress management and demonstrate how data-driven approaches can support academic institutions in improving student well-being.

As per Stress Dependency:

The analysis conducted in this project indicates that students' ability to manage stress is influenced by multiple lifestyle and academic factors. Among these, **sleep patterns and time management emerged as the most significant contributors to effective stress management**. Students who reported adequate sleep and structured daily routines demonstrated a higher ability to handle stress. In contrast, **academic workload and perceived academic pressure showed a negative impact**, with increased pressure leading to reduced stress-handling capacity. Additionally, overall lifestyle balance, including regular breaks and consistency in daily habits, played a supportive role in stress management. These findings highlight that stress is a **multidimensional phenomenon**, dependent on the combined effect of sleep quality, time management, academic demands, and lifestyle balance rather than a single isolated factor.

FUTURE SCOPE

While this project successfully applies predictive analytics techniques to analyze student lifestyle and stress management using survey data, there are several opportunities for further enhancement and expansion. One potential area for future work is the **collection of longitudinal data**. Instead of a one-time survey, data could be collected over multiple academic terms to analyze changes in student behavior and stress levels over time. This would allow for trend analysis and more accurate prediction of long-term stress patterns.

Another important extension of this study is the incorporation of **advanced machine learning and deep learning models**. Techniques such as Artificial Neural Networks (ANN), Multi-Layer Perceptrons (MLP), and ensemble methods like Random Forests and Gradient Boosting could be implemented to improve predictive accuracy. These models are capable of capturing complex, non-linear relationships that may not be fully addressed by traditional regression and logistic regression approaches.

The scope of the project can also be expanded by integrating **additional data sources**, such as academic performance records, attendance data, or activity tracking information (with proper consent and ethical approval). Combining survey data with behavioral or performance-based metrics would provide a more comprehensive understanding of the factors influencing student stress and well-being.

Furthermore, future research may explore **real-time stress monitoring systems** by incorporating wearable device data, mobile application inputs, or periodic micro-surveys. Such systems could enable early detection of stress-related issues and support timely interventions by academic counselors or mental health professionals. Predictive models developed in this project could serve as the foundation for building intelligent recommendation systems for personalized student support.

Finally, the study can be extended to a **larger and more diverse population** by collecting data from multiple institutions, academic disciplines, or geographic regions. This would improve the generalizability of results and allow comparative analysis across different student groups. Overall, these future directions highlight the potential of predictive analytics to evolve into a powerful decision-support tool for enhancing student well-being and academic success.

REFERENCES

1. Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.
2. Müller, A. C., & Guido, S. (2017). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media.
3. Weidman, S. (2019). *Deep Learning from Scratch: Building with Python from First Principles*. O'Reilly Media.
4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in Python*. Springer.
5. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
6. Scikit-learn Documentation. Retrieved from https://scikit-learn.org
7. McKinney, W. (2018). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media.
8. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 90–95.
9. Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters*, 31(8), 651–666.
10. Google Forms Documentation. Retrieved from https://support.google.com/docs

LINK

GOOGLE FORM LINK:

https://forms.gle/vF5Nt3r17Q2azMYm6

GITHUB LINK:

https://github.com/vaanshgarg/Student-Lifestyle-Well-being-Survey

LINKEDIN:

https://www.linkedin.com/posts/vaanshgarg_datascience-predictiveanalytics-studentlife-activity-7407108165475061760-bWZF?utm_source=social_share_send&utm_medium=android_app&rcm=ACoAAEfCzgMBIS8vIV83S5Db_vfk7rcHsdHuywo&utm_campaign=whatsapp

GOOGLE DRIVE LINK:
https://drive.google.com/drive/folders/1dky_vCccgTIB5aqlLHfNMqgKu-lfP9dC?usp=drive_link