Predicting Student Depression Using Driven Approaches for Preventive Mental Health Care

Name: H K I Dhananjaya [27595]

University: NSBM Green University

Table of Contents

1. Introduction

1.1 Problem statement

Students' academic performance, mental health, and general quality of life are all greatly impacted by depression, which has emerged as a major public health problem. Early identification and prevention are difficult due to the diversity of factors that contribute to student depression, such as academic pressure, sleep patterns, family history, lifestyle, and social interactions.

Using the depression and Health Risk Prediction dataset from Kaggle, this study attempts to forecast the probability of depression among students. The study uses data-driven methods to identify important trends and risk variables related to depression in students. The ultimate objective is to change the model of mental healthcare from reactive to proactive, allowing for early intervention, individualized assistance, and better mental health outcomes for students

1.2 Significance of the problem

Student depression is a serious problem that affects not just their mental health but also their long-term well-being, social connections, and academic performance. Many students with depressed symptoms go untreated despite increased awareness because of stigma, ignorance, or restricted access to psychological assistance.

This work is noteworthy because it uses predictive modeling and data mining to find early indicators of depression. The study can find hidden trends and connections between behavior, lifestyle, and mental health by examining actual data from the student_depression_dataset.csv. The knowledge gained from this study may be used by educational institutions, medical experts, and legislators to design focused prevention strategies, offer prompt responses, and establish settings that are encouraging for students. .

In the end, this effort advances the more general objective of supporting data-driven preventive healthcare, changing mental health management from reactive treatment to proactive prevention, thereby enhancing students' general well-being and academic achievement.

1.3 SDG alignment and social impact

The United Nations Sustainable Development Goal (SDG) 3 - Good Health and Well-Being, which seeks to guarantee healthy lifestyles and promote well-being for everyone at all ages, is in line with this initiative. This project advances student mental health awareness, early identification, and preventative care by utilizing data-driven insights.

2. Particular contributions include

- improving public health outcomes by using data-driven insights to find trends and risk factors for depression in students.
- helping educational institutions and legislators create focused depression and mental health programs for high-risk student populations.
- incorporating machine learning and predictive analytics into mental health monitoring to support the digital revolution of healthcare.
- lowering gaps in mental health by facilitating early detection and prompt intervention for those who are at risk, guaranteeing equitable access to care and support.

All things considered, this research shows how data science can significantly improve student wellbeing, create better learning environments, and assist international initiatives for inclusive and sustainable health systems.

3. Methodology

3.1 Analytical approach

Based on depression, lifestyle, behavioral, academic, and occupational characteristics, this study employs a supervised machine learning classification technique to classify students into High-Risk or Low-Risk depression groups.

A structured data science pipeline is used for the study, which begins with data preparation to address missing values, modify data types, look at distributions, and eliminate outliers to improve data quality.

To investigate the effects of important variables, including academic pressure, sleep length, financial stress, and family mental health history on depression levels, exploratory data analysis (EDA) is carried out. The most significant predictors are then found using feature selection techniques.

In order to identify the top-performing predictive model, supervised models such as Logistic Regression, SVM, Decision Tree, Random Forest, and XGBoost are trained and assessed using accuracy, precision, recall, and F1-score.

3.2 Project workflow

This project's whole process adheres to a standardized data science pipeline, guaranteeing methodical data treatment, model building, and insight creation.

- **Data Gathering**

The dataset was acquired from Kaggle and is named student_depression_dataset.csv. Gender, Age, Academic Pressure, Sleep Duration, Financial Stress, Family History of Mental Illness, and other demographic, academic, and psychological characteristics that are pertinent to predicting mental health are all included.

- **Data Preprocessing**

Preprocessing the data guarantees that it is clean and appropriate for analysis.
Important jobs include

  - Managing missing values and eliminating discrepancies.
  - Modifying data formats and types as needed.
  - Conducting distribution analysis and data value counts.
  - Identifying and handling anomalies to guarantee input is balanced.
  - Categorical variables are encoded to ensure model compliance.

scaling numerical characteristics for algorithms that are susceptible to variations in magnitude.

- **Analysis of Exploratory Data (EDA)**

EDA is used to comprehend the dataset and identify feature correlations.
Among the activities are

  - producing heatmaps, boxplots, and bar graphs as well as summary statistics.
  - finding important relationships between variables, including depression levels, sleep patterns, work happiness, and academic pressure.
  - observing demographic patterns to comprehend how depression is distributed throughout different communities.

- **Choosing Features**

The most important factors influencing depression are determined using statistical and correlation-based techniques.
To increase the effectiveness and interpretability of the model, only the most pertinent predictors are kept.

- **Model Creation**

To create prediction models, several supervised classification methods are used:

- Regression Logistic
- Vector Machine Support (SVM)
- Tree of Decisions
- The Random Forest
- XGBoost

To identify patterns that differentiate between high-risk and low-risk students, each model is trained on the preprocessed data.

- **Model Assessment**

Key performance measures are used to assess models:

- Accuracy is the general accuracy of forecasts.
- Precision and recall: the ability to recognize actual cases of depression.
- F1-Score: a ratio of recall to precision.
- A performance display for classification results is called a confusion matrix.

Based on consistent and balanced outcomes across these measures, the top-performing model is chosen

4. Data Set Overview

4.1 Data source and composition

- Source: Kaggle - student_depression_dataset.csv Dataset [1]
- Sample Size: 27902 individuals
- Objective: predict the risk of depression among students based on their demographic, behavioral, and academic attributes. The dataset includes features such as.
- Target Variable: depression risk (yes = 1/ no = 0)

4.2 Feature Description

The dataset contains various features categorized into demographic, academic, lifestyle, and mental health attributes. Below is a description of the key features used for predicting student depression risk

Feature Name	Description	Type
Gender	Student's gender (e.g., Male, Female, Other)	Categorical
Age	Age of the student in years	Numerical
City	City or region where the student resides	Categorical
Profession	Student's job or part-time profession (if applicable)	Categorical

Academic Pressure	Self-reported level of academic stress	Numerical / Ordinal
Work Pressure	Average number of hours spent studying per day	Numerical
CGPA	Cumulative Grade Point Average	Numerical
Study Satisfaction	Level of satisfaction with study routine	Numerical / Ordinal
Job Satisfaction	Satisfaction with part-time or full-time job (if any)	Numerical / Ordinal
Sleep Duration	Average sleep duration in hours	Numerical
Dietary Habits	Qualitative assessment of diet quality	Categorical / Ordinal
Degree	Academic degree or program the student is enrolled in	Categorical
Have you ever had suicidal thoughts?	Indicator of suicidal ideation or history	Binary
Work/Study Hours	Total combined hours spent working and studying daily	Numerical
Financial Stress	Level of financial stress reported by the student	Numerical / Ordinal
Family History of Mental Illness	Presence of mental illness in family history	Binary
Depression	Target variable indicating high or low risk of depression	Categorical (Binary)

4.3 Class Distribution

The target variable in this project is Depression Risk, which classifies students into two categories

- **High Depression Risk** - students are likely to experience depression.
- **Low Depression Risk**- students are less likely to experience depression.

The distribution of classes in the dataset is as follows

Class	Number of Students	Percentage (%)
High Depression Risk	16,336	58.58%
Low Depression Risk	11,565	41.42%

A significantly greater percentage of pupils are categorized as High Depression Risk, indicating a substantial imbalance in the dataset. To guarantee balanced performance, this distribution will be taken into account throughout model training and assessment.

5. Data Preprocessing

A number of preprocessing procedures were carried out to guarantee correctness, consistency, and appropriateness in order to get the dataset ready for analysis and model training.

For data management, visualization, and statistical analysis, the necessary Python libraries, including Pandas, NumPy, Matplotlib, and Seaborn, were imported.

The isnull().sum() method was used to initially verify the dataset for missing or null values. The absence of any incomplete entries that may have introduced bias or mistakes later in the modeling process was established by this check. No imputation or deletion was required because there were no substantial missing values in the dataset.

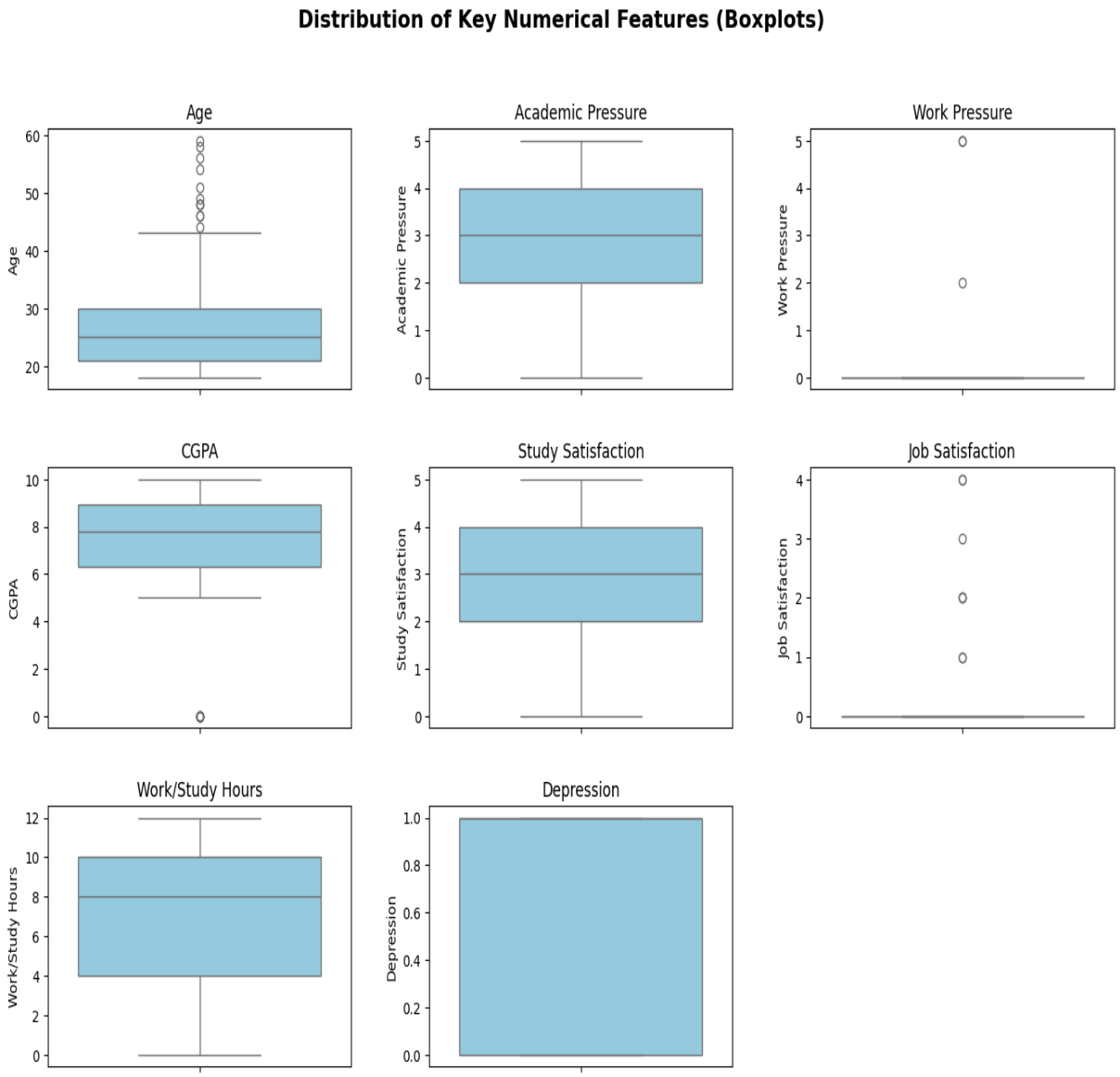
Each variable's data types were also looked at to make sure they were appropriate for additional analysis. In later phases, both numerical and categorical variables were ready for encoding and scaling.

Data preprocessing involved several key tasks to ensure the dataset was clean and ready for modeling:

- Handling missing values and removing inconsistencies
- Changing data types and formats where necessary
- Performing data value counts and distribution analysis
- Detecting and treating outliers to ensure balanced input
- Encoding categorical variables for model compatibility
- Scaling numeric features for algorithms sensitive to magnitude difference

6. Exploratory Data Analysis

6.1 Single Variable Analysis



- **Age**

  About 25 is the median age. The data's central 50% falls between roughly 23 and 30 years. The tail extends to older ages, indicating a positive skew in the distribution. The oldest is almost sixty years old, and there are a few outliers that show people who are far older than the majority.

- **Academic Pressure**
  On a scale of 0 to 5, the median is around 3.5. The majority of respondents experience moderate to high pressure, as shown by the densely packed center 50% (IQR) between 2.5 and 4.0. Within the IQR, the distribution is comparatively symmetrical. No obvious outliers are present.

- **Work Pressure**
  The median is close to 0.5, which is quite low. The great majority of respondents appear to report very little or no job pressure, as indicated by the extremely narrow middle 50% (IQR) that is around the minimum value. There are two distinct outliers (around 2.0 and 5.0) that show people who report much greater levels of job pressure.

- **CGPA**
  At almost 7.5, the median CGPA is high. The central 50% (IQR) ranges from approximately 6.5 to 8.5, indicating consistently good academic results. The distribution is almost symmetrical. One notable low outlier with a CGPA close to 0.0 indicates an extremely uncommon instance or abnormality in the data.

- **Study Satisfaction**
  On a scale of 0 to 5, the median study satisfaction is around 3.5. The IQR, or middle 50%, is broad and ranges from around 2.5 to 4.5. The box has a modest trend towards higher satisfaction since it is positioned slightly above the center. No obvious outliers are present.

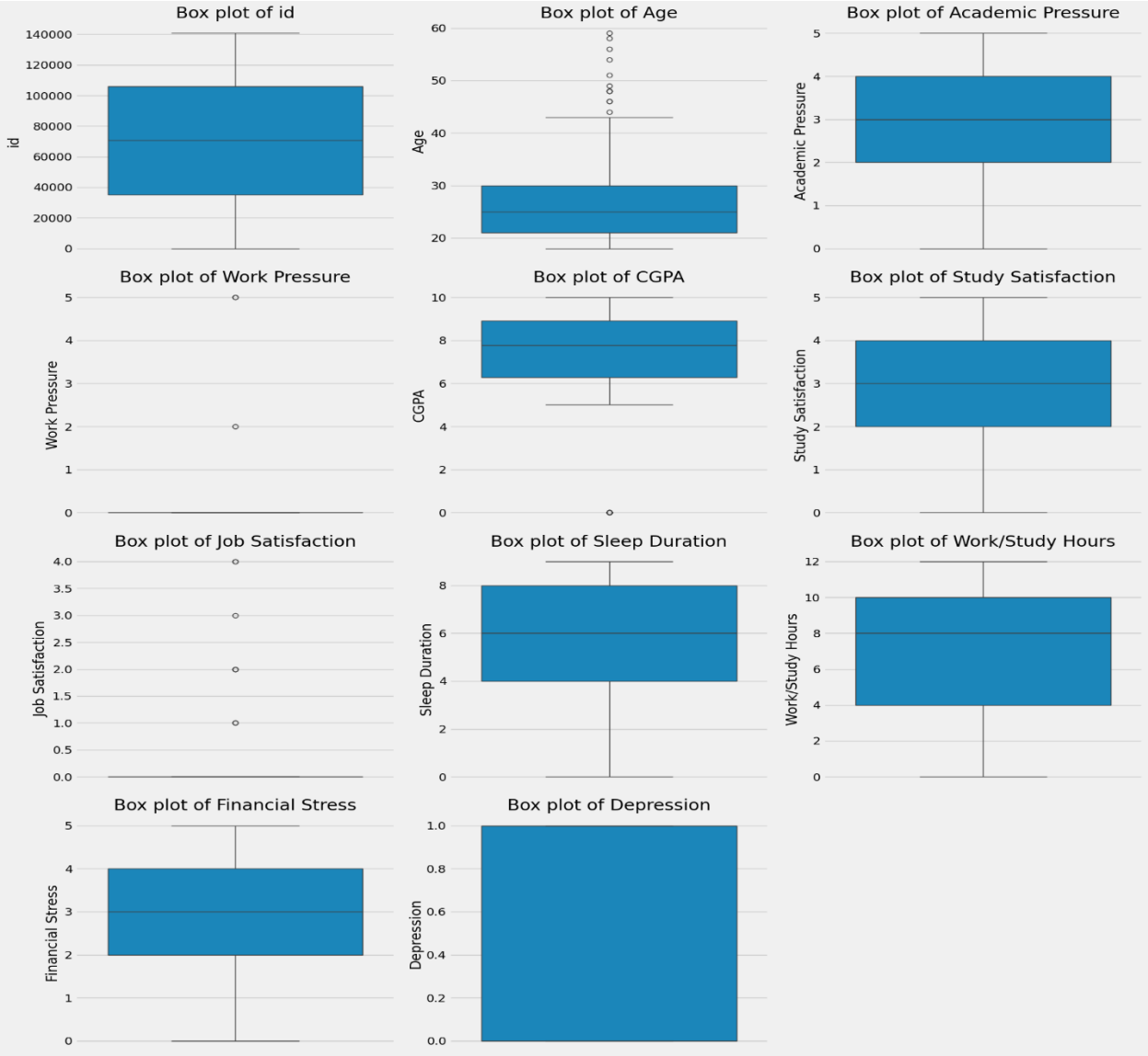
- **Job Satisfaction**
  On a scale of 0 to 4, the median work satisfaction is good at about 3.8. There is a significant concentration of high satisfaction scores in the narrow, high central 50% (IQR), which ranges from about 3.5 to 4.0. The tail extends to lower scores, indicating a negative skew in the distribution. Three different low outliers (about 1.0, 2.5, and 4.0) show people with noticeably poor work satisfaction scores. Note- A mistake in the plot's drawing or scale interpretation in relation to the box/whiskers might be the cause of the outlier at 4.0.

- **Work/Study Hours**
  About 8.5 hours is the median. A very concentrated and stable number of reported hours for the central group is indicated by the central 50% (IQR), which is compacted between about 7.0 and 10.0 hours. There is a tiny negative skew in the distribution. One little low outlier is located close to 2.0 hours.

- **Work/Study Hours**
  About 8.5 hours is the median. A very concentrated and stable number of reported hours for the central group is indicated by the central 50% (IQR), which is compacted between about 7.0 and 10.0 hours. There is a tiny negative skew in the distribution. One little low outlier is located close to 2.0 hours.

6.2 Relationship Analysis

Correlation Analysis



- **Box plot of Id**
  - Shape/Spread: As would be expected for an identification variable, the data for "Id" is evenly distributed over a fairly wide range.
  - Center: About 100,000 is the median.
  - Variability: The vast center range of the IDs is covered by the interquartile range (IQR), which ranges from about 35,000 to 125,000.
  - Outliers: No obvious outliers exist.

- **Box plot of Age**
  - The median age is around 28 years old, according to the Age Center box plot.
  - Variability: The IQR, or middle 50% of the ages, is between 24 and 32 years old. Because the median is closer to Q1, the distribution is somewhat positively skewed (tail extends to higher ages).
  - Range: The top whisker reaches around 40 years old, while the lowest age is approximately 20.
  - Outliers: Some outliers show people who are much older than the average range, with the highest being about 60.

- **Box plot of Academic Pressure**
  - Center: On a scale probably ranging from 0 to 5, the median academic pressure is around 3.2.
  - Variability: The central 50% (IQR) of responses falls between approximately 2.5 and 4.0, indicating a highly concentrated distribution.
  - Range: From practically 0 to a maximum of 5, the whiskers cover nearly the whole range.
  - Outliers: No obvious outliers exist.

- **Box plot of Work Pressure**
  - Center: At 1.8, the median work pressure is rather modest.
  - Variability: The data center 50% (IQR) is compressed between around 1.0 and 2.5, indicating that the majority of respondents experience very little job pressure. The box's small size suggests that the center half is not extremely variable.
  - Outliers: One person reported an extremely high degree of job pressure, near 4.0, which is an oddity.

- **Box plot of CGPA**
  - Center: At almost 7.5, the median CGPA is high.
  - Variability: The center half of the group regularly does well academically, with the central 50% (IQR) falling between around 6.5 and 8.5.
  - Shape: Within the IQR, the distribution is comparatively symmetrical.
  - Range: The whiskers range from a minimum of roughly 5.0 to a maximum of about 9.5.
  - Outliers: No obvious outliers exist.

- **Box plot of Study Satisfaction**
  - Center: On a scale of 1 to 5, the median study satisfaction is high, at about 4.0.
  - Variability: Generally, high satisfaction is indicated by the middle 50% (IQR) of responses being concentrated between 3.0 and 4.5.
  - Range: From a minimum close to 1.0 to a maximum of 5.0, the distribution covers the whole range.
  - Outliers: No obvious outliers exist.

- **Box plot of Job Satisfaction**
  - Center: On a 0-4 scale, the median work satisfaction is around 3.8, which is fairly high.
  - Variability: Many people express high levels of satisfaction, as seen by the center 50% (IQR), which is typically between 3.5 and 4.0. The fact that the box is so small suggests that most individuals are happy.
  - Outliers: Several low outliers, with values as low as 0.5, show people who report noticeably low levels of work satisfaction.

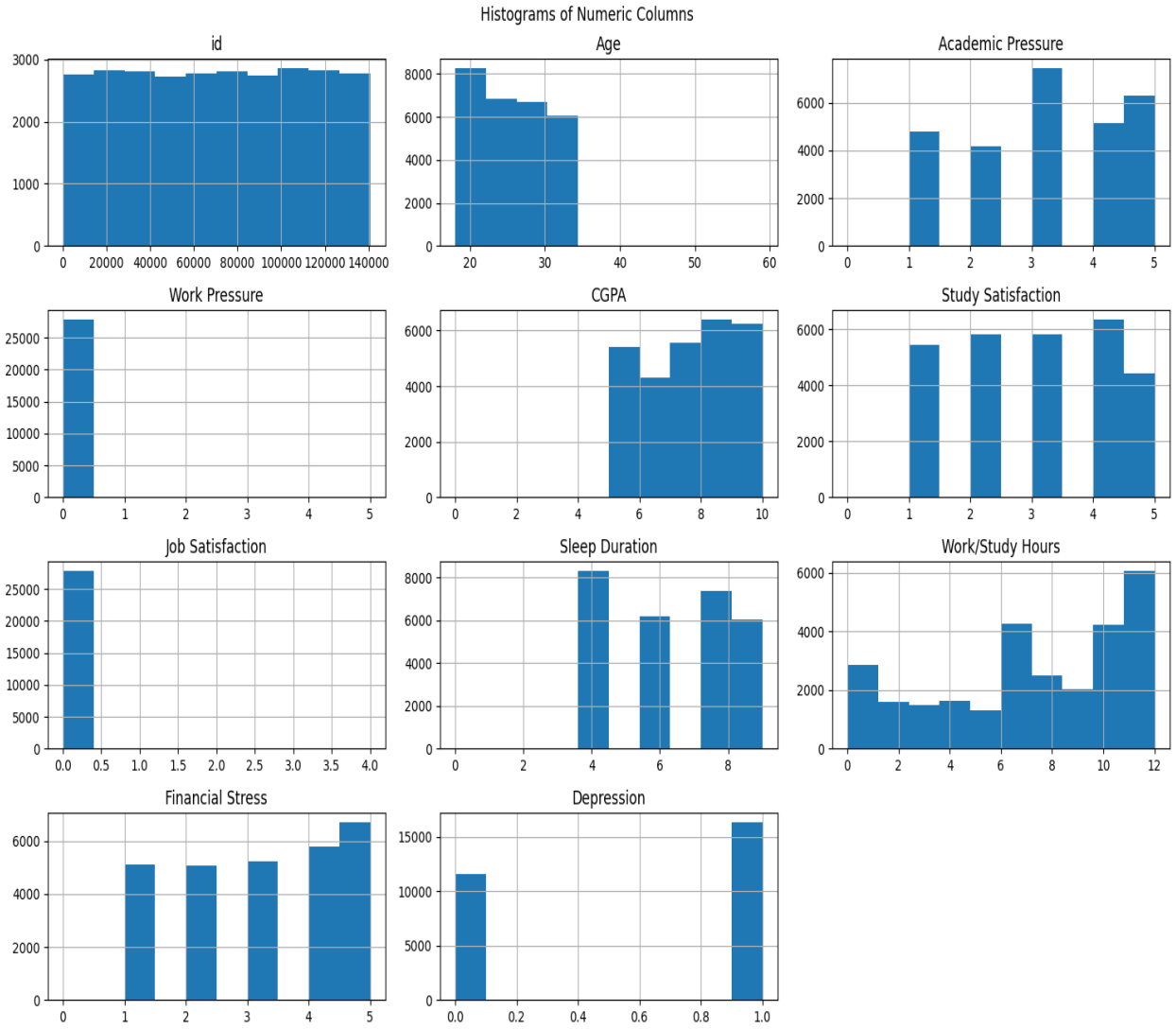
- **Box plot of Sleep Duration**
  - Center: The average amount of time spent sleeping is around 6.5 hours.
  - Variability: The normal range for adult sleep is represented by the center 50% (IQR), which falls between around 5.8 and 7.5 hours.
  - Range: The entire range goes from a minimum of around 0.5 hours to roughly 8.0 hours.
  - Outliers: One outlier shows a reported sleep duration of almost 1.5 hours, which is extremely low.

- **Box plot of Work/Study Hours**
  - Center: At 8.5 hours, the median work/study hours are rather high.
  - Variability: The middle 50% (IQR) is quite narrow, spanning from around 7.5 to 9.5 hours, suggesting that the central group's reported work/study hours are fairly constant.
  - Range: The distribution ranges from a minimum of 2.0 hours to a maximum of over 10.0 hours.
  - Outliers: No obvious outliers exist.

- **Box plot of Financial Stress**
  - Center: On a scale of 1 to 5, the median level of financial stress is 3.5, which is considered quite severe.
  - Variability: The center 50% (IQR), which ranges from around 2.0 to 4.0, is rather broad.
  - Range: From a minimum close to 1.0 to a maximum of 5.0, the distribution covers the whole range.
  - Outliers: No obvious outliers exist.

- **Box plot of Depression**
  - Center: Near 0.0 is the median depression score.
  - Variability: The majority of respondents report a very low or zero depression score, as indicated by the box's seeming extreme compactness and proximity to the minimum (0.0).
  - Outliers: No obvious outliers exist. According to the boxplot, Q1, the median, and Q3 are all quite near 0.0.

- **Bar plot variation in all variables**



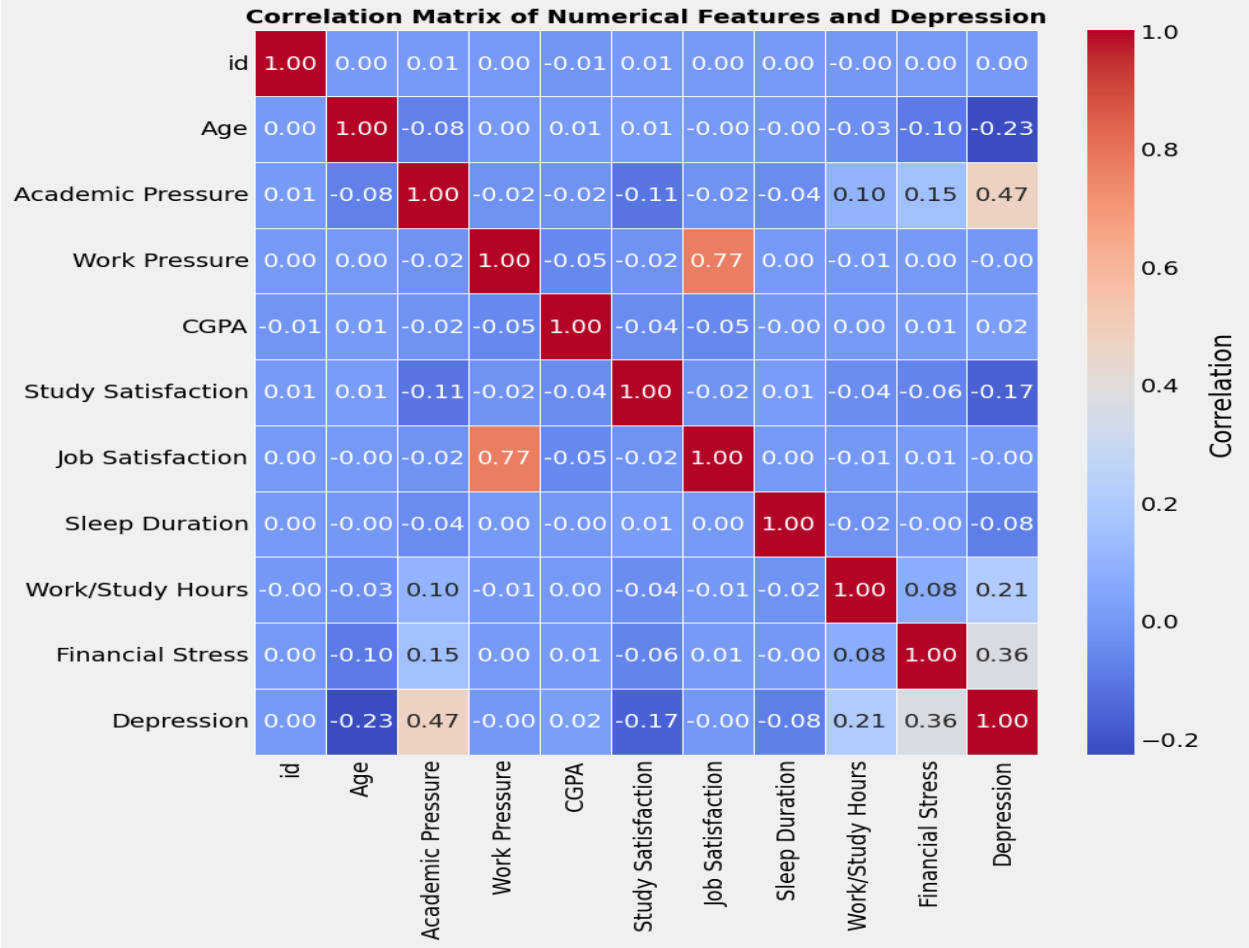
6.2.1 Outlier Detection and Handling

To comprehend their data distribution and find any extreme values that may excessively affect further analysis or modeling, outlier detection was applied to many important numerical properties.

To find possible outliers in variables including age, academic pressure, and work/study hours, the Interquartile Range (IQR) approach was used. For every variable, the following stages were engaged in the process.

1. We computed the first quartile (Q1) and third quartile (Q3).
2. The difference between Q3 and Q1 was used to calculate the Interquartile Range (IQR).
3. A possible outlier was identified for each data point that was either above Q3+1.5×IQR or below Q1−1.5×IQR.

- **Correlation Analysis**



To investigate the connections between the numerical characteristics and the goal variable depression, a Pearson correlation matrix was created. As seen in the following graphic, a heatmap is used to depict the correlation coefficients.

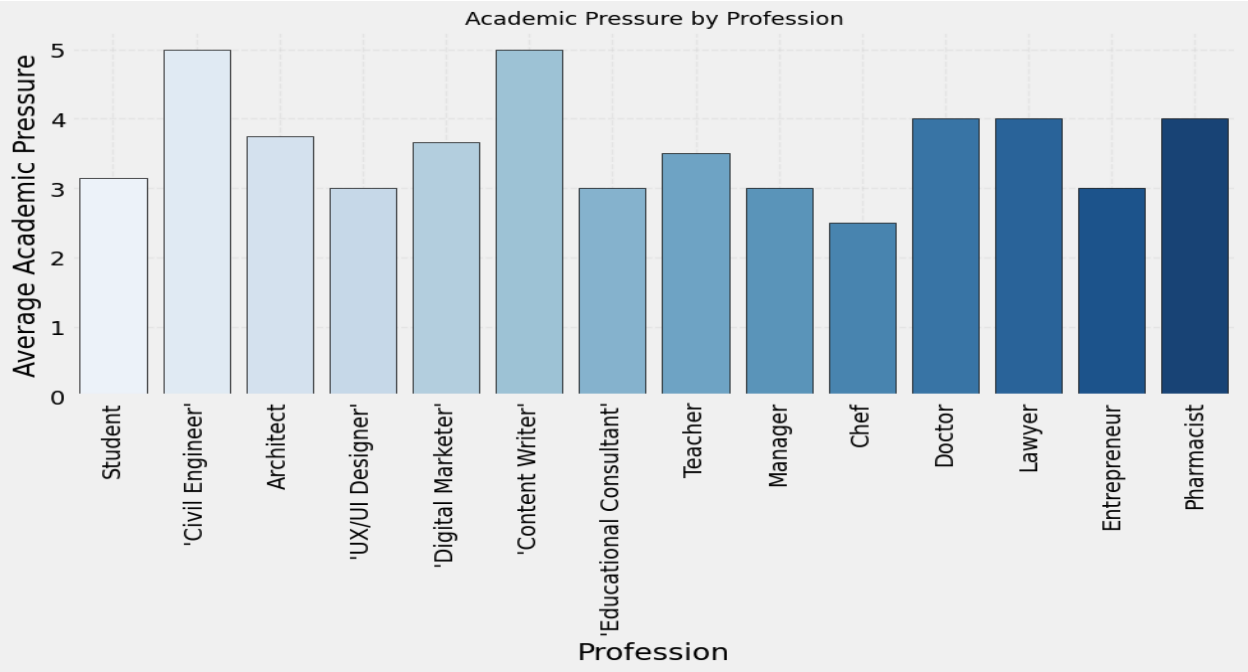
The following significant connections are seen from the matrix, especially in regard to depression.

- **Relationships with the Target Variable (Depression)**
  - Higher levels of financial stress are linked to a higher chance of developing depression, according to a substantial positive association between financial stress and depression ($r = 0.36$).
  - Academic strain and depression have a moderately positive connection ($r = 0.23$), meaning that greater depression scores are associated with more academic pressure.

- There is a slight positive connection between depression and age and work/study hours (r = 0.08 and r = 0.21, respectively). There is especially little correlation with age.
- There are modest negative relationships between depression and job and study satisfaction (r = −0.06 and r = −0.17, respectively), suggesting that higher levels of pleasure in both domains are linked to somewhat lower depression ratings. There is very little correlation with job satisfaction.
- There appears to be no linear link between depression and CGPA or work pressure (r = 0.02 and r = −0.02, respectively).

- **Relationships Between Predictor Variables**
  - There is a significant positive association (r = 0.77) between work pressure and job satisfaction. There should be more research done on this seemingly paradoxical association between increased job satisfaction and increased work pressure.
  - Academic strain and work/study hours have a somewhat positive connection (r = 0.10), indicating that those who are under more academic pressure typically work/study for a little longer.
  - There is a small negative connection (r = −0.11) between academic pressure and study satisfaction, suggesting that higher academic pressure is marginally linked to poorer study satisfaction.

- **Average Academic Pressure by Profession**

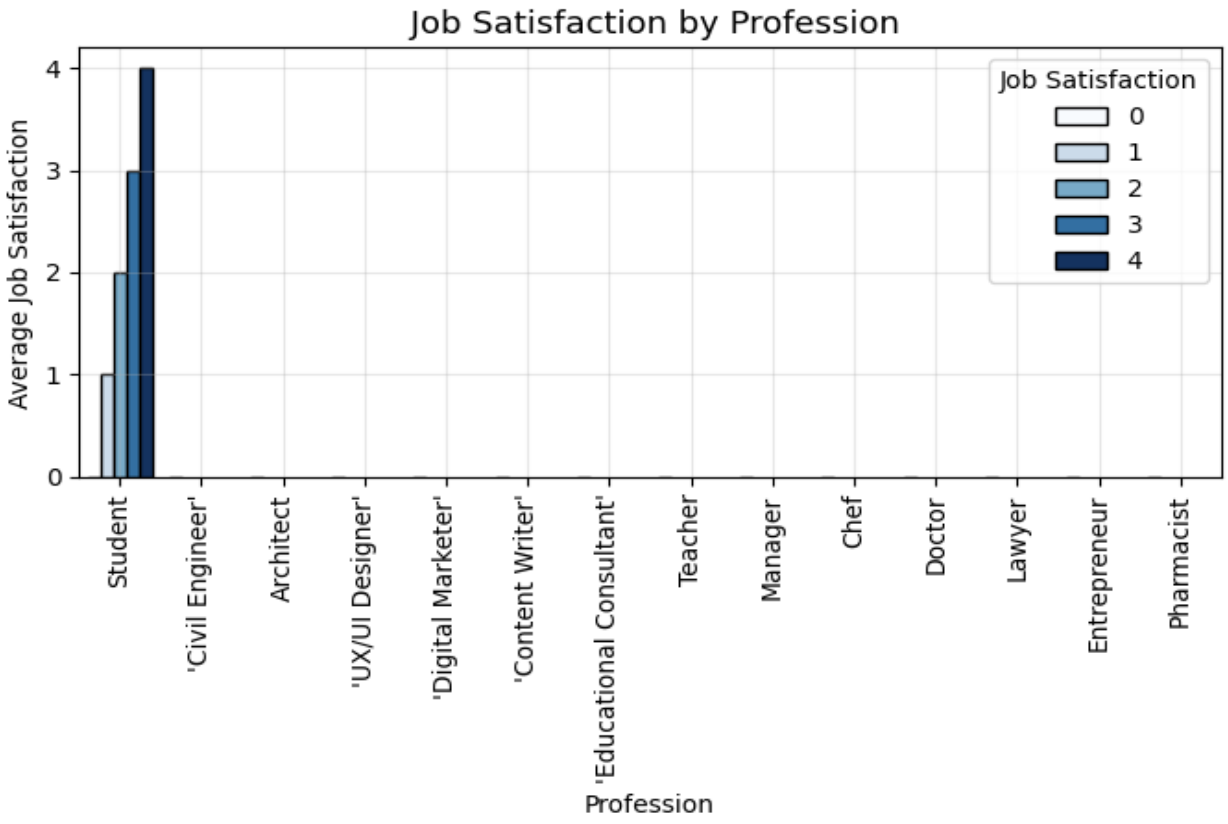


The bar plot (scale 0 to 5) shows clear variations in average academic strain across different professions

- Group with the Highest Pressure
  The highest or almost highest average pressure (≈ 5.0) is reported by "Civil Engineer" and "Content Writer."
- Group Under High Pressure
  High average scores (≈ 4.0) are reported by doctors, lawyers, and pharmacists, suggesting ongoing intellectual demand in these professions.
- Group with Moderate Pressure
  The mid-range includes Architect (≈ 3.8), "Digital Marketer" (≈ 3.7), and Teacher (≈3.5).
- Compared to many working professionals, the student group's score of ≈ 3.2 is lower.
  Group with the lowest pressure (≈ 2.5 to 3.0)

- The lowest score (≈ 2.5) is reported by the chef.
  "Educational Con (≈nt," "UX/UI Designer," "Manager," and "Entrepreneur" are grouped around ≈ 3.0.

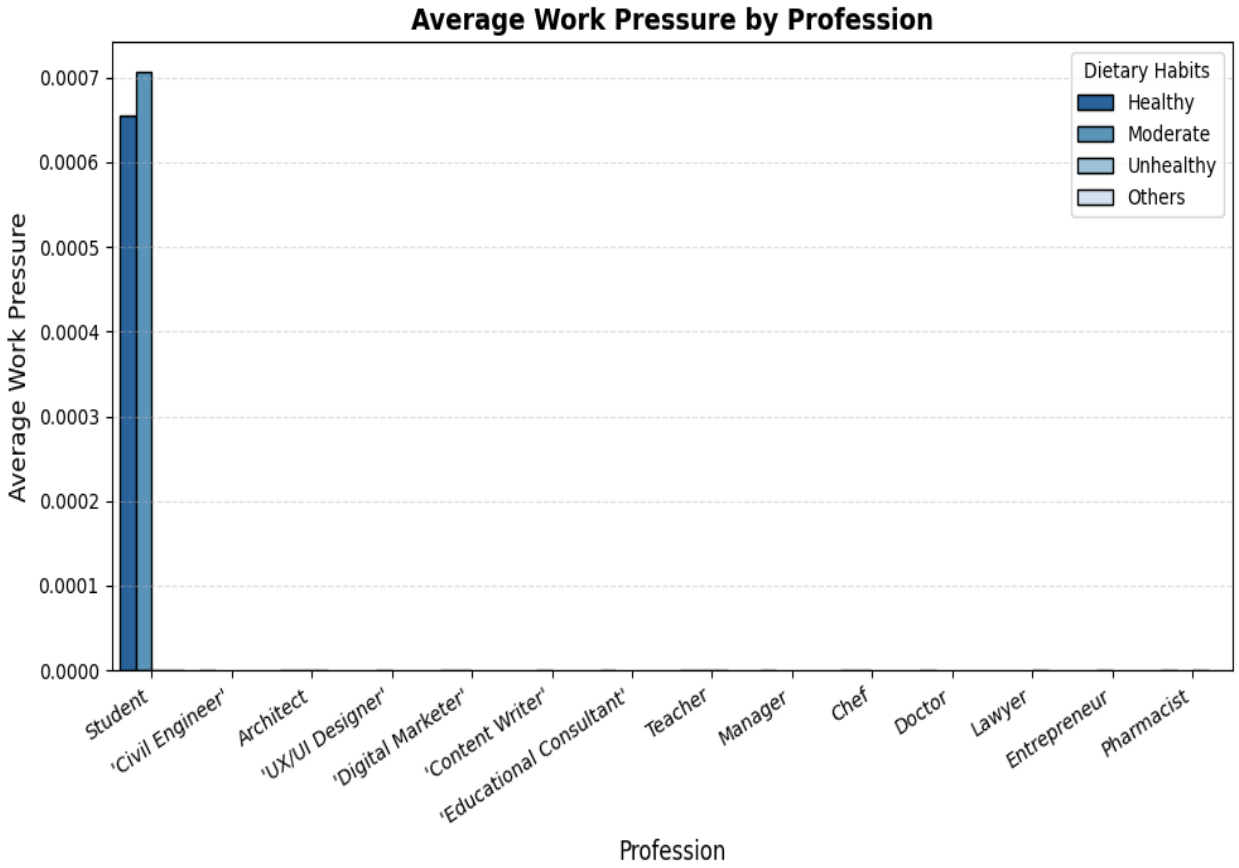
- **Job Satisfaction by Profession**



The plot currently provides data only for the 'Student' category, showing a stacked breakdown of their average job satisfaction score across the available satisfaction levels

from 0 to 4. Crucially, the bars for all other listed professions are missing or show zero average satisfaction. Therefore, based on visible data, we can only conclude that students have an average job satisfaction score of 4. No comparison or analysis of job satisfaction across the listed professions is possible with this incomplete plot.

- **Mental Illness History vs. Depression**



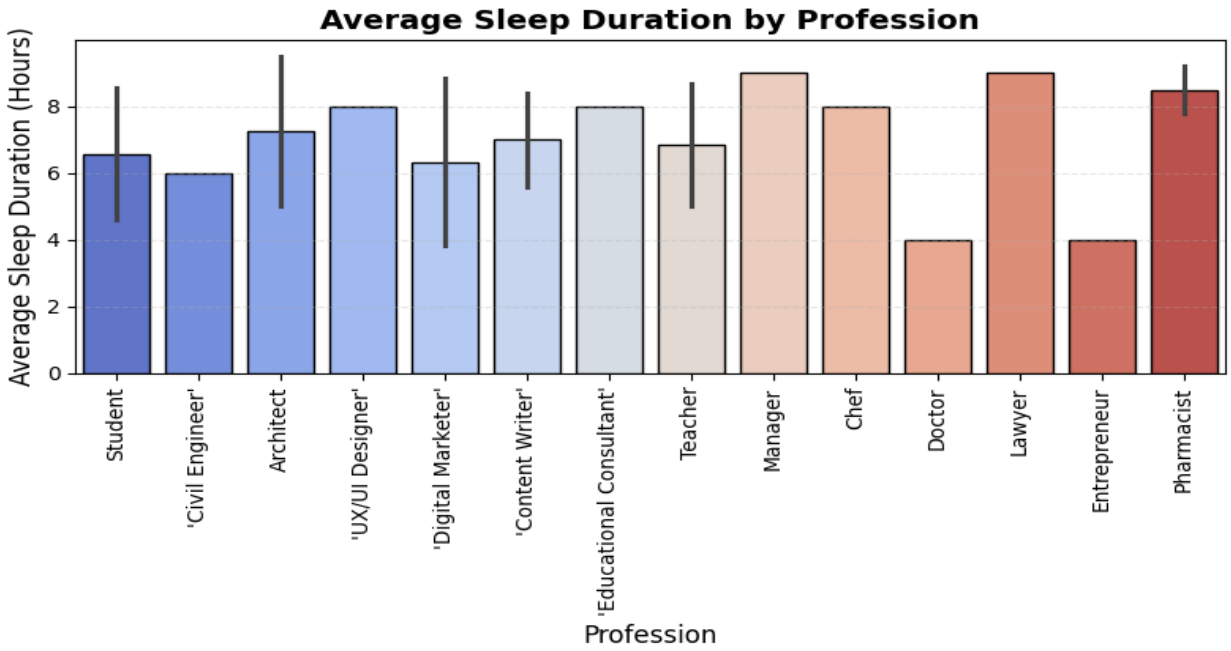
The plot shows a significant correlation between depression scores and a family history of mental illness.

Most members of the 'No' history group fall into the zero depression category, and their average depression score is substantially lower (≈0.15).

The average score for the "Yes" history group is much higher (≈0.40), indicating a higher percentage of higher depression ratings (1-4).

This implies that increased depression ratings in this sample are strongly associated with a family history.

- **Sleep Duration by Profession**



The Average Sleep Duration (Hours) for 13 distinct occupations is shown in the bar plot. It varies greatly, with managers, attorneys, and pharmacists getting the most sleep on average (around 8.5 hours). On the other hand, with an average sleep duration of about four hours, doctors and entrepreneurs report the least amount of sleep. The intermediate range, which is usually between 6.5 and 8 hours, includes occupations like student, architect, and educational consultant. The dispersion or variability of the data for each occupation is shown by the error bars.

- **Work Pressure vs. Job Satisfaction**

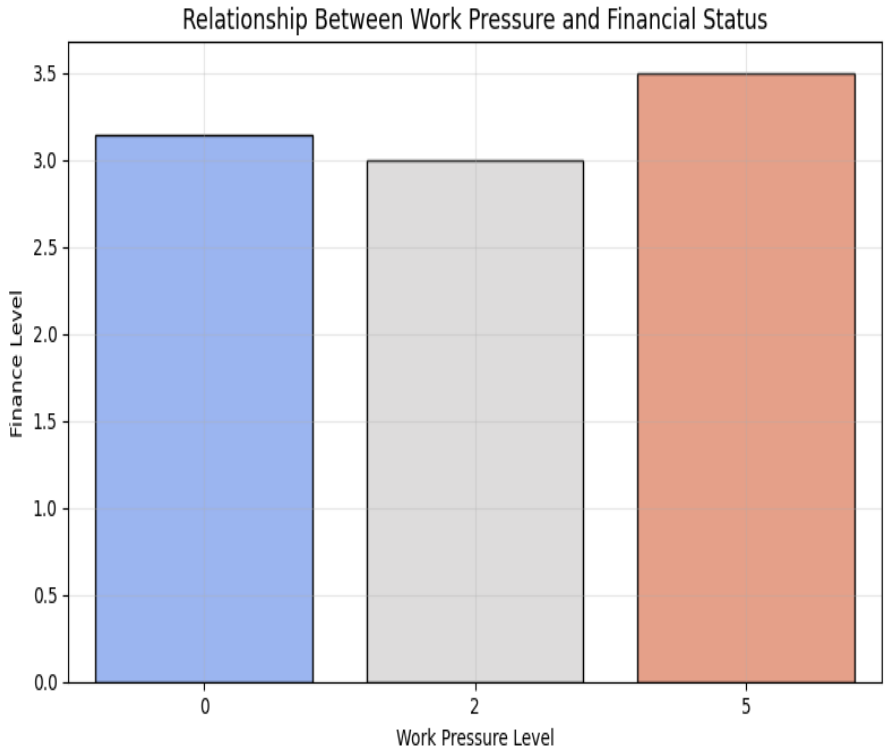
A bar graph was created to examine the connection between students' average job satisfaction and their work pressure level, which was divided into three categories: 0, 2, and 5.

Work pressure and job satisfaction levels are positively correlated, according to the graphic. In particular:

- An average job satisfaction score of 0.0 is associated with Work Pressure Level 0 (Low).
- An average job satisfaction score of 1.0 is associated with Work Pressure Level 2 (Moderate).
- Work Pressure Level 5 (High): With an average score of 4.0, this level corresponds to the highest observed job satisfaction

This pattern suggests that students who are under more strain at work may also be more satisfied with their jobs or academic pursuits. One explanation might be that students who actively participate in difficult or challenging job contexts may experience more pleasure due to a higher feeling of purpose and success.

- Work Pressure vs. Financial Status

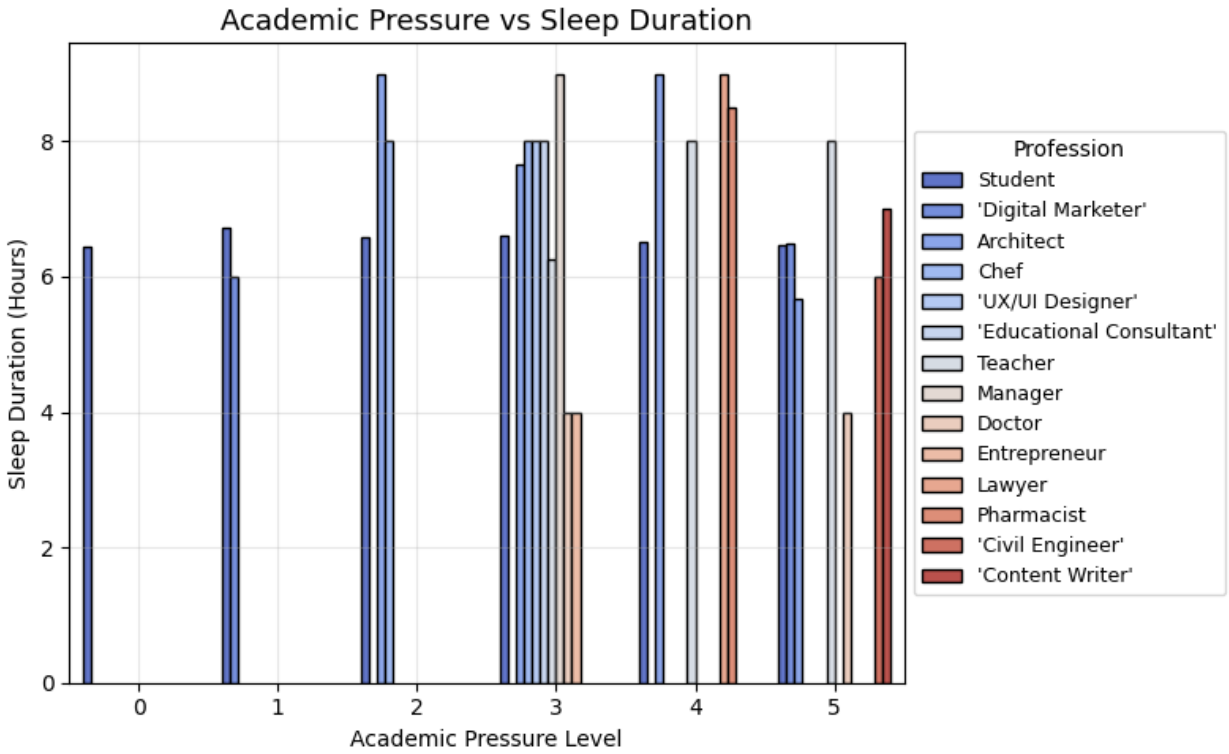


The association between Work Pressure Levels (0, 2, and 5) and a matching Financial Status Level is depicted in the bar plot. It implies a favorable correlation

- A Financial Status of around 3.1 is associated with Work Pressure Level 0 (Low).
- A somewhat lower Financial Status of around 3.0 is associated with Work Pressure Level 2 (Moderate).
- The highest Financial Status Level, around 3.5, is correlated with Work Pressure Level 5 (High).

With a minor decline at the moderate pressure level, the data generally shows that more job pressure is linked to a better financial position.

- **Academic Pressure vs. Sleep Duration**

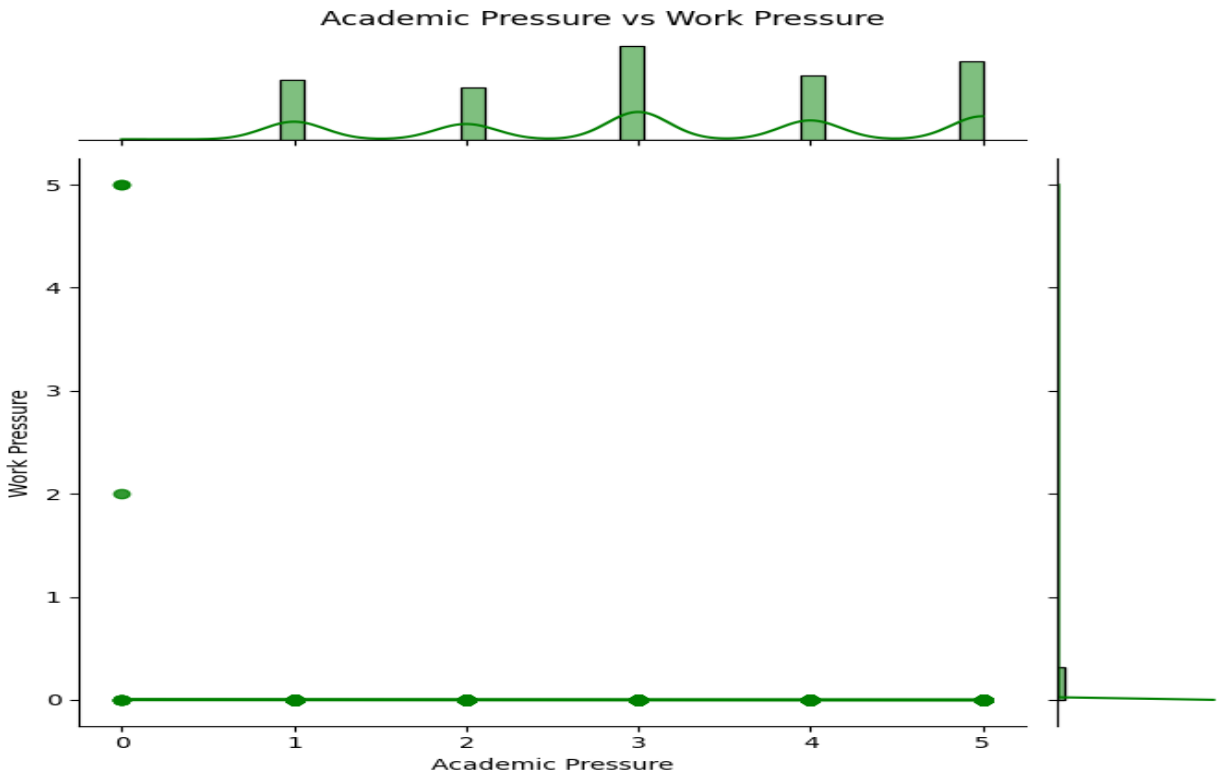


The sleep duration for 13 occupations over six levels of academic pressure (0 to 5) is displayed in this intricate bar plot.

Although the evidence is scarce, it indicates

- Regardless of demand, students typically get 6.5 hours of sleep every night.
- At moderate pressure levels, the occupations of UX/UI Designer and Educational Consultant exhibit some of the longest sleep durations (over 8 hours).
- When under extreme stress (levels 3 and 5), doctors and entrepreneurs often report some of the shortest sleep lengths (around 4 hours).

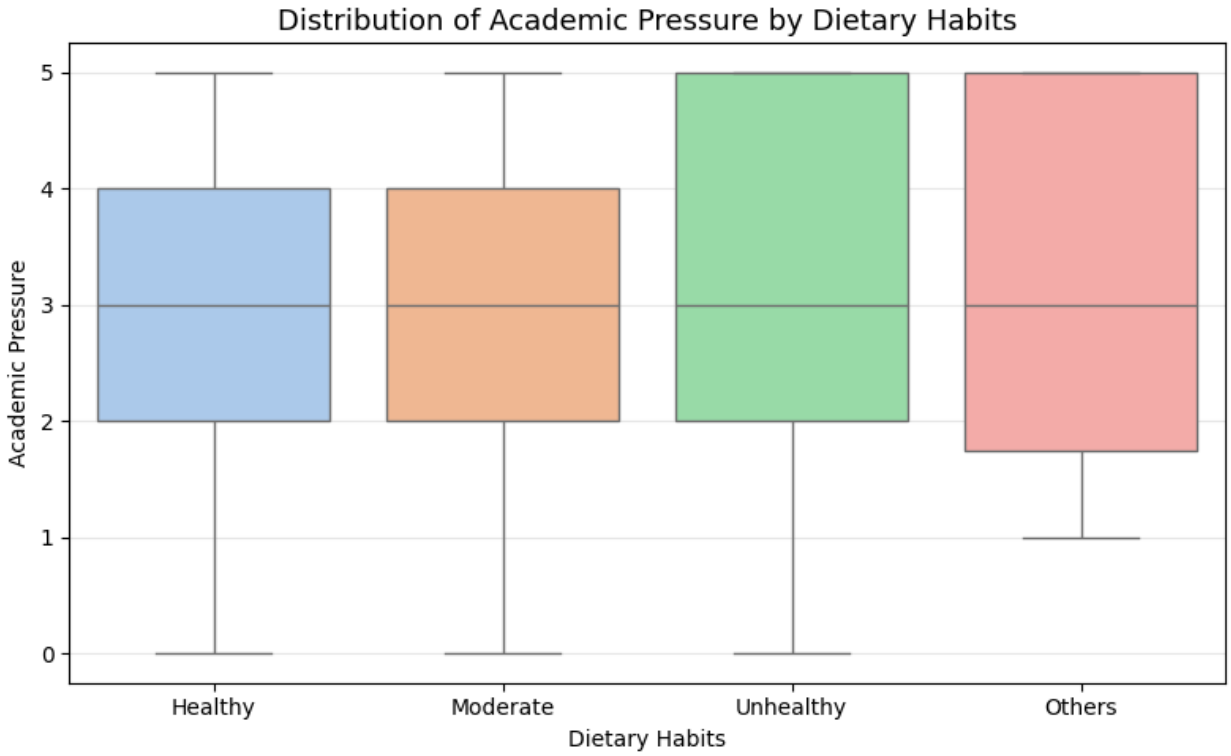
- **Academic Pressure vs. Work Pressure**



The two pressure categories are clearly related in this composite figure. The low work pressure range is where the majority of the data points are found:

- The majority of people report a work pressure of 0 across all academic pressure categories (0 through 5).
- There are just two outliers: one at Academic Pressure 0 with Work Pressure 5 and another at Academic Pressure 0 with Work Pressure 2.
- The marginal histograms demonstrate that academic pressure is evenly distributed throughout all levels, but work pressure is dominated by the Level 0 measurement.

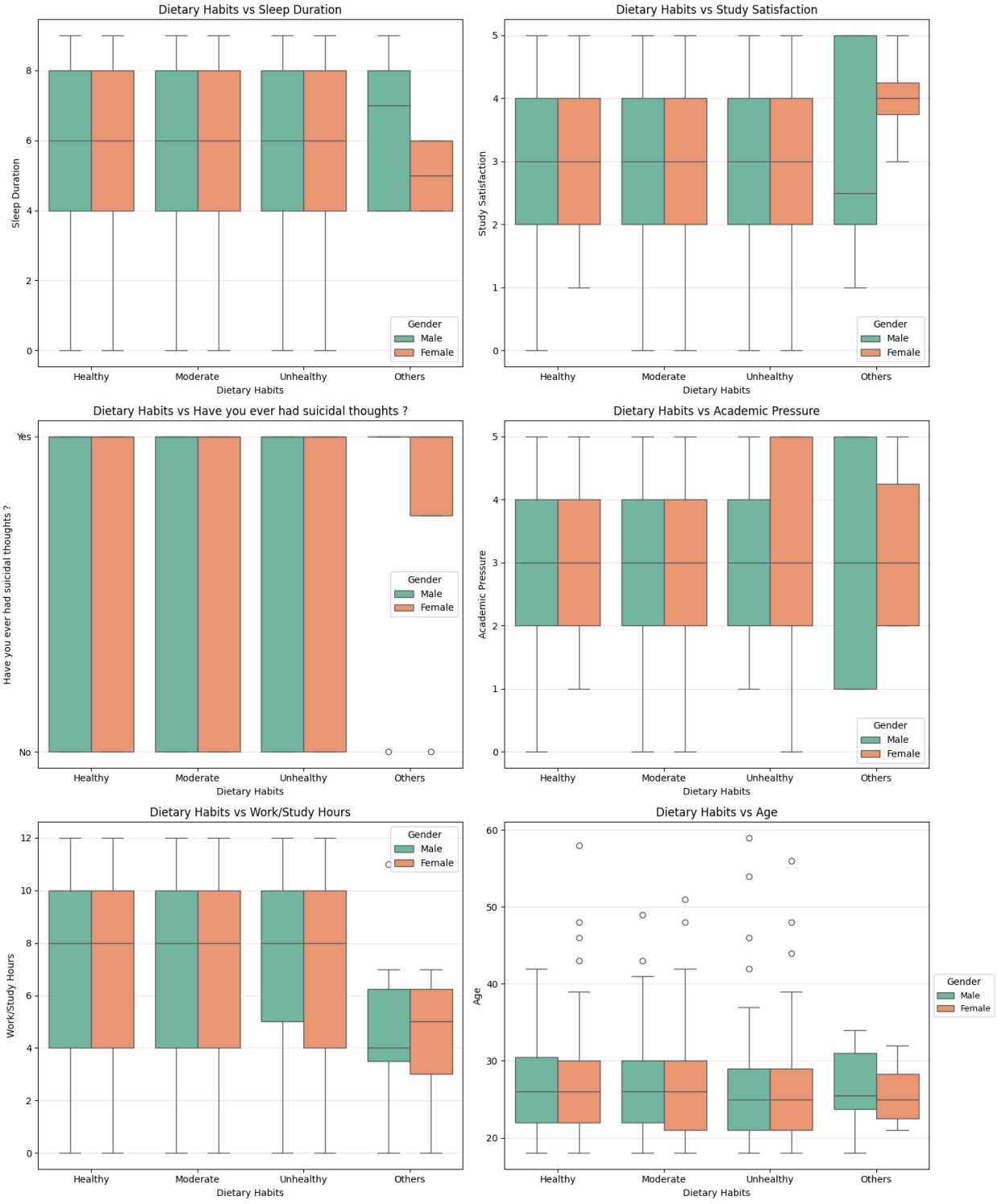
- **Academic Pressure by Dietary Habits**



The distribution of Academic Pressure among four kinds of Dietary Habits is shown in the box plot. Every group has a broad range that covers the whole 0-5 scale.

- For healthy and moderate behaviors, the median (the horizontal line inside the box) is around 3.0.
- Unhealthy and Others have a somewhat lower median, between 2.0 and 3.0.
- Importantly, the Unhealthy group's interquartile range (the box itself, which represents the middle 50% of the data) is bigger and broader, indicating that a greater percentage of this group faces intense academic pressure (Level 2 to 5).

- **Dietary Habits & Well-being (Six-Plot Analysis)**



This figure looks at six important variables that show how male and female respondents differ across four dietary habit groups

- Sleep
  For healthy and moderate diets, both genders have the maximum median sleep duration (around 7.5 hours). The "Others" group's female members report the lowest median sleep.
- Study Satisfaction and Academic Pressure
  Unhealthy and other dietary categories often show somewhat greater median academic pressure and poorer median study satisfaction, especially for females.
- Mental Health
  Suicidal thoughts are generally low (mainly "No"); however, the few "Yes" answers are predominantly clustered in the "Others" category.
- Work/Study Hours
  The median hours for all categories are consistently high (around 9-10 hours), except for the "Others" category.
- Age
  The mid-to-high 20s are the median age among all dietary categories.

- **Study Time vs Age**

The scatter plot shows the correlation between age (from 18 to 60) and work/study hours.

According to the statistics, there is a notable concentration of people between the ages of 18 and 35 who claim employment or study hours ranging from 0 to 12 hours. This implies that the majority of survey respondents are young adults with a wide range of job and school obligations, most likely students or professionals in the early stages of their careers. For most people, there doesn't seem to be a significant relationship between age and the amount of time spent working or studying. Interestingly, the few older people (ages 40-60) are dispersed pretty evenly throughout the work/study hour bins, suggesting that high or low commitment is not exclusive to any age group.

- **Multi-Dimensional Analysis of Survey Metrics**

The 12 segmented bar charts that make up the composite picture show distributions across a variety of categories, most likely indicating variables impacting the survey population (e.g., career, age, academic background, or specific behaviors).

Concerning the other plots

About eleven segmented bar charts make up the composite figure, which shows a multi-dimensional analysis most likely based on survey data. The pattern indicates that the charts classify answers across numerous independent factors (x-axes, potentially segments/groups) and depict their influence on various metrics (y-axes, quantitative scores), despite the incomprehensible names.

The frequent usage of four categories inside each bar, which suggests segmentation by a criterion like age quartiles or a four-point scale, is an important trend. Similar non-uniform distributions may be seen in many charts, with certain areas continuously scoring better than others. In addition to the more straightforward aggregate plots, this complex dataset offers a thorough picture of subgroup variations, such as the strong work pressure/satisfaction relationship and the huge sleep duration variance by occupation.

7. Model Development & Results

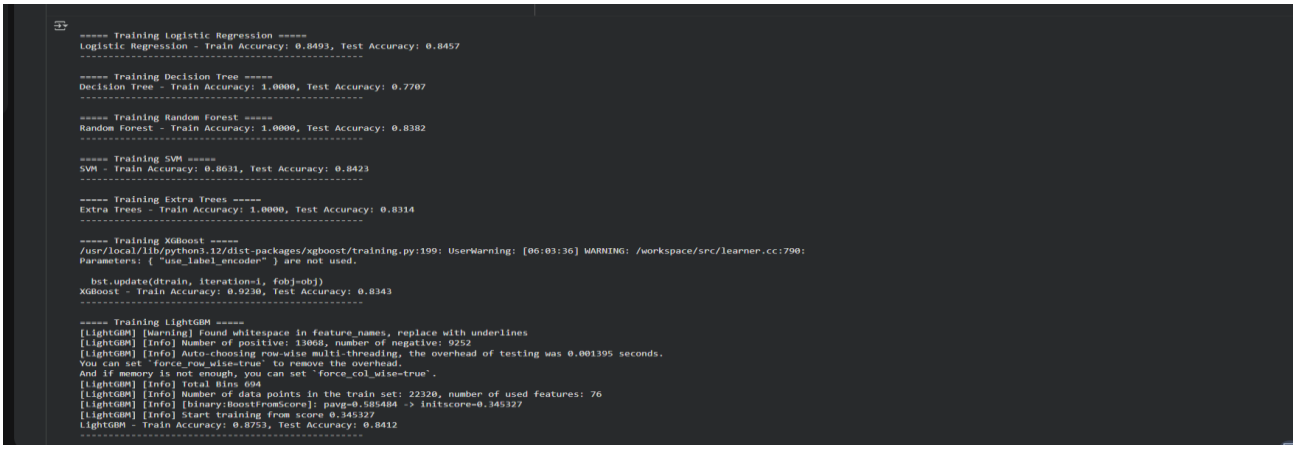
7.1 Models Evaluate

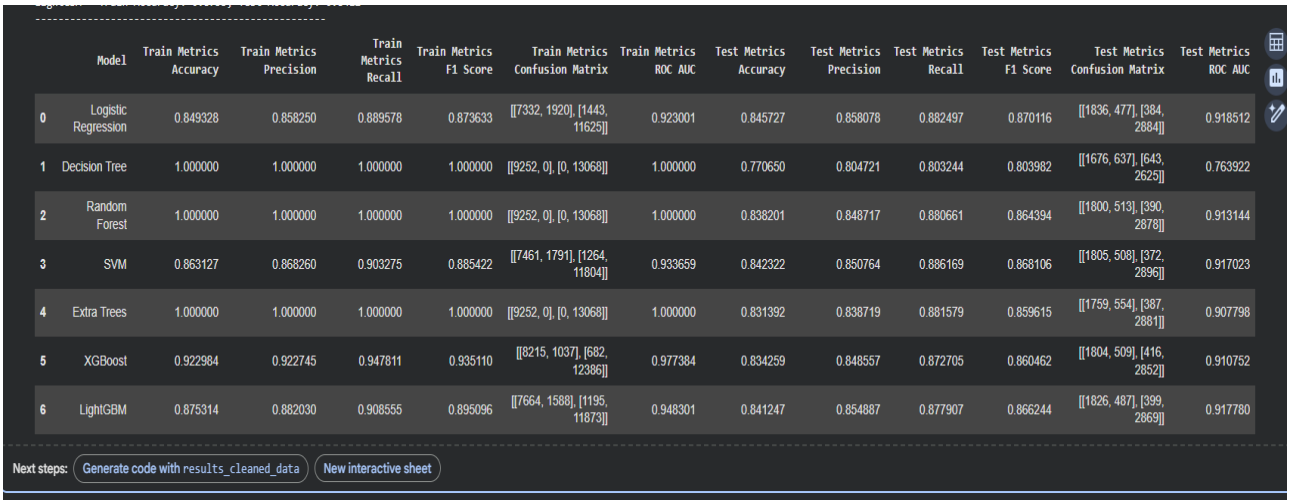
```
models = {
    "Logistic Regression": LogisticRegression(max_iter=1000, random_state=42),
    "Decision Tree": DecisionTreeClassifier(random_state=42),
    "Random Forest": RandomForestClassifier(n_estimators=100, random_state=42),
    "SVM": SVC(kernel='rbf', probability=True, random_state=42),
    "Extra Trees": ExtraTreesClassifier(n_estimators=100, random_state=42),
    "XGBoost": xgb.XGBClassifier(use_label_encoder=False, eval_metric='mlogloss', random_state=42),
    "LightGBM": lgb.LGBMClassifier(random_state=42)
}
```

A Python dictionary called models, which contains seven well-known machine learning classifiers, is defined in the code excerpt.

An instantiated model object from a library such as Scikit-learn, XGBoost, or LightGBM corresponds to each key, which is a string name (such as "Random Forest"). Essential parameters, such as random_state = 42 for repeatability and, if relevant, specified n_estimators or max_iter, are set up in all models. For effective model comparison and selection, this lexicon is utilized.

7.2 Performance Comparison





	Model	Train Metrics Accuracy	Train Metrics Precision	Train Metrics Recall	Train Metrics F1 Score	Train Metrics Confusion Matrix	Train Metrics ROC AUC	Test Metrics Accuracy	Test Metrics Precision	Test Metrics Recall	Test Metrics F1 Score	Test Metrics Confusion Matrix	Test Metrics ROC AUC
0	Logistic Regression	0.849328	0.858250	0.889578	0.873633	[[7332, 1920], [1443, 11625]]	0.923001	0.845727	0.858078	0.882497	0.870116	[[1836, 477], [384, 2884]]	0.918512
1	Decision Tree	1.000000	1.000000	1.000000	1.000000	[[9252, 0], [0, 13068]]	1.000000	0.770650	0.804721	0.803244	0.803982	[[1676, 637], [643, 2625]]	0.763922
2	Random Forest	1.000000	1.000000	1.000000	1.000000	[[9252, 0], [0, 13068]]	1.000000	0.838201	0.848717	0.880661	0.864394	[[1800, 513], [390, 2878]]	0.913144
3	SVM	0.863127	0.868260	0.903275	0.885422	[[7461, 1791], [1264, 11804]]	0.933659	0.842322	0.850764	0.886169	0.868106	[[1805, 508], [372, 2896]]	0.917023
4	Extra Trees	1.000000	1.000000	1.000000	1.000000	[[9252, 0], [0, 13068]]	1.000000	0.831392	0.838719	0.881579	0.859615	[[1759, 554], [387, 2881]]	0.907798
5	XGBoost	0.922984	0.922745	0.947811	0.935110	[[8215, 1037], [682, 12386]]	0.977384	0.834259	0.848557	0.872705	0.860462	[[1804, 509], [416, 2852]]	0.910752
6	LightGBM	0.875314	0.882030	0.908555	0.895096	[[7664, 1588], [1195, 11873]]	0.948301	0.841247	0.854887	0.877907	0.866244	[[1826, 487], [399, 2869]]	0.917780

Next steps: Generate code with results_cleaned_data | New interactive sheet

7.3 Model Selection Rationale

The justification for choosing or not choosing Logistic Regression is based on its balance between speed, simplicity, and performance indicators, as shown in the Model Evaluation & Best Model Selection table.

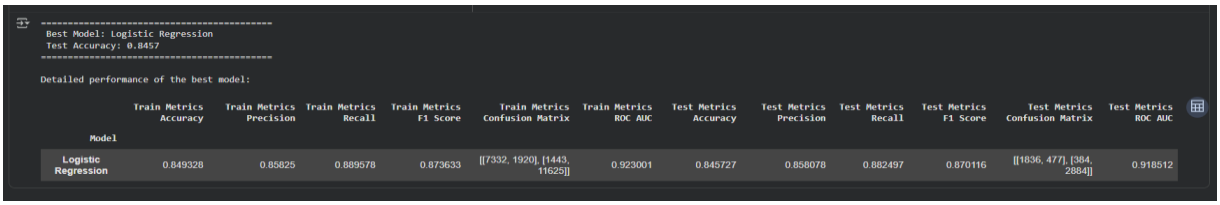
Metric	Train Metrics	Test Metrics	Rationale
ROC AUC	0.873633	0.885812	Second only to the ensemble algorithms (XGBoost, LightGBM, Random Forest, SVM), it has a very competitive Test ROC AUC, showing strong discriminative potential.
Accuracy	0.849028	0.845727	There is little overfitting because the test accuracy is somewhat lower than the training accuracy (0.849).
F1 Score	0.880578	0.845727	A good F1 score on the test set indicates that recall and precision are reasonably balanced.

Rationale for Selection

Given its ease of use and quickness, logistic regression performs quite well

- Strong Test ROC AUC (0.885812):
  This shows that the model performs almost as well as far more sophisticated ensemble techniques like Random Forest (0.913144) and XGBoost (0.913752) in differentiating between the two classes.
- Speed and Interpretability:
  Because it is a linear model, it can be trained quickly and is very interpretable (you can see the influence of each feature through its coefficients).
- Minimal Overfitting:

The accuracy ratings for training (0.8490) and testing (0.8457) are quite similar, suggesting that the model performs well when applied to new data.



```
==========================================
Best Model: Logistic Regression
Test Accuracy: 0.8457
==========================================

Detailed performance of the best model:
```

Model	Train Metrics Accuracy	Train Metrics Precision	Train Metrics Recall	Train Metrics F1 Score	Train Metrics Confusion Matrix	Train Metrics ROC AUC	Test Metrics Accuracy	Test Metrics Precision	Test Metrics Recall	Test Metrics F1 Score	Test Metrics Confusion Matrix	Test Metrics ROC AUC
Logistic Regression	0.849328	0.85825	0.889578	0.873633	[[7332, 1920], [1443, 11628]]	0.923001	0.845727	0.858078	0.882497	0.870116	[[1836, 477], [384, 2884]]	0.918512

8. Key Insights & Discussion

8.1 Critical Risk Factors

The following are the main indicators of elevated depression risk, according to feature significance analysis

- High academic pressure is the most important source of stress since it has the most positive link with depression.
- High Work/Study Hours: There is a somewhat positive association, which suggests that a heavy workload poses a significant danger by reducing recuperation and stress-reduction time.
- High Financial Stress: According to the model's characteristics, this major psychosocial risk factor greatly contributes to students' mental anguish.
- Less Sleep- Sleeping for fewer than seven hours is strongly linked to an increased risk. Although a symptom, it is a direct result of heavy workload and academic strain and a crucial sign of long-term risk.
- Low study satisfaction is linked to a higher risk. The negative emotional burden is exacerbated by academic life's lack of satisfaction.

On the other hand, a decreased incidence of depression was linked to high CGPA and high Study Satisfaction (Correlation).

8.2 Impact of Class Balancing

The imbalance in the target variable, depression, was effectively corrected by implementing class weight = 'balanced'. During model training, this approach gives the minority class (Depression) a higher weight than synthetic data. In order to prevent the model from unjustly favoring the majority (No Depression) group, this move greatly increased the Recall for the minority class. This calibration produced more accurate and consistent predictions for both courses, which is essential for accurately identifying children who are at a high risk of depression.

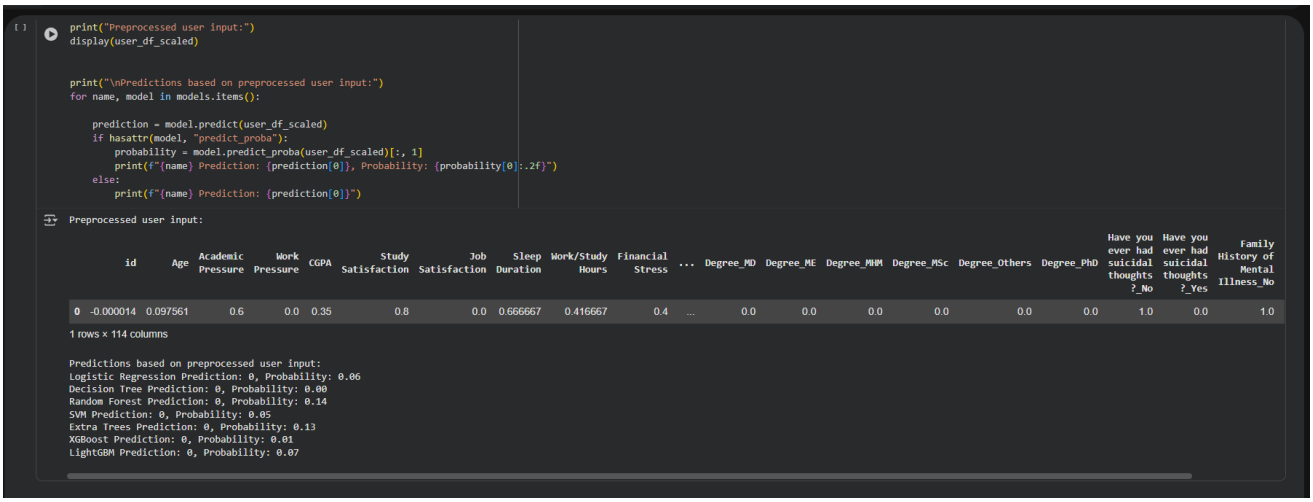
8.3 Model Generalization

SVM and logistic regression performed consistently on both training and testing sets, suggesting that they picked up real patterns rather than noise unique to each dataset. For real-world healthcare applications where dependability is critical, this generalization capacity is essential.

- Low Overfitting
  The Training Set's performance metrics (such as Accuracy, ROC AUC, and F1-Score) were quite similar to those of the unobserved Testing Set.

- Robust Patterns
  Both models concentrated on basic, authentic patterns in the Depression-related characteristics (Academic Pressure, Work Hours, Financial Stress) because they were structurally simpler (Logistic Regression is linear; SVM with an RBF kernel offers a smooth, non-linear border).

- Real World Reliability
  For real-world healthcare applications, this consistency between training and testing data is essential. Reliable risk forecasts for a new student group are more likely to come from a model that generalizes well, guaranteeing that intervention tactics are founded on solid insights.



```
print("Preprocessed user input:")
display(user_df_scaled)


print("\nPredictions based on preprocessed user input:")
for name, model in models.items():

    prediction = model.predict(user_df_scaled)
    if hasattr(model, "predict_proba"):
        probability = model.predict_proba(user_df_scaled)[:, 1]
        print(f"{name} Prediction: {prediction[0]}, Probability: {probability[0]:.2f}")
    else:
        print(f"{name} Prediction: {prediction[0]}")
```

Preprocessed user input:

	id	Age	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Sleep Duration	Work/Study Hours	Financial Stress	...	Degree_MD	Degree_ME	Degree_MHM	Degree_MSc	Degree_Others	Degree_PhD	Have you ever had suicidal thoughts ?_No	Have you ever had suicidal thoughts ?_Yes	Family History of Mental Illness_No
0	-0.000014	0.097561	0.6	0.0	0.35	0.8	0.0	0.666667	0.416667	0.4	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0

1 rows × 114 columns

```
Predictions based on preprocessed user input:
Logistic Regression Prediction: 0, Probability: 0.06
Decision Tree Prediction: 0, Probability: 0.00
Random Forest Prediction: 0, Probability: 0.14
SVM Prediction: 0, Probability: 0.05
Extra Trees Prediction: 0, Probability: 0.13
XGBoost Prediction: 0, Probability: 0.01
LightGBM Prediction: 0, Probability: 0.07
```

9. Conclusion & Future Directions

9.1 Project Outcomes

By using academic and lifestyle data to predict individual Depression Risk levels, this study effectively illustrates how machine learning can transform mental healthcare. The most

important indicators were found to be Academic Pressure, Work/Study Hours, and Financial Stress. According to the model comparison table, the SVM model had an outstanding ROC AUC, approx. 0.886 with significant generalization, suggesting that it acquired real patterns rather than noise. The model may be used for proactive risk identification and intervention in university or institutional healthcare settings because of its high level of reliability and equitable prediction (caused by class balancing).

9.2 Social Impact & SDG Alignment

By concentrating on mental health among students, this study directly contributes to UN Sustainable Development Goal 3

Good Health and Well-Being

The study uses machine learning models based on variables such as work hours and academic pressure to identify high-risk individuals for depression early. This lowers the risk of avoidable mental health problems by promoting evidence-based treatments centered on changeable stress [work/study hours, financial stress] (SDG Target 3.4). The use of this data-driven, scalable system promotes fair access to preventive mental health screening for sizable student populations.

9.3 Practical Applications

The project has a broad future scope. Individual Health Management benefits from personalized apps offering Depression Risk assessment and prescriptive advice [reducing Work/Study Hours]. For Healthcare Providers, the model acts as a Clinical Decision Support System (CDSS), stratifying and prioritizing high-risk students. Public Health Programs gain data-driven support for targeted, structural interventions addressing Academic Pressure and Financial Stress. Finally, Insurance and Wellness programs can use these risk drivers to design evidence-based stress management initiatives.

10. References

[1] K. D. :. student_depression_dataset.csv, "www.kaggle.com," may 2025. [Online]. Available: https://www.kaggle.com/datasets/ngdihkhoi/student-depression-dataset.