Predicting Depression in Students based on Social and Behavioral Factors

Project Report - Data 606 Statistical Methods in Data Science

FEBRUARY 20, 2025

Marcelino Rodriguez, Rajvir Kaur, Nyadual Makuach, Asiah Zibrila
Statistical Methods in Data Science

Contents

Tables and Figures:

1. Introduction

Everyone feels sad, moody, or low on energy from time to time. However, for some people, these feelings persist intensely for weeks, months, or even years. Depression is more than just feeling down; it's a serious condition that affects both physical and mental health. According to the World Health Organization (WHO), one in eight people worldwide suffers from a mental health disorder (Liu & Wang, 2024). Among these, depression is one of the most common, affecting over 280 million people. Despite its prevalence, the way depression is treated varies across different communities. This makes it essential to raise awareness, especially in communities where mental health issues are not taken seriously. As Marks (2022) highlights, individuals from BIPOC communities are more likely to be diagnosed with depression, yet they are less likely to seek help. Ignoring mental health struggles does not make them disappear. In fact, untreated depression can lead to more severe consequences, including chronic pain, substance abuse, self-harm, and even suicide. Addressing mental health concerns promptly is crucial to improving overall well-being and quality of life

Students going through transitional life stages are more likely to experience mental health challenges, including depression. Research shows that the number of students struggling with depression and anxiety is rising (Grineski et al., 2024; Xiao et al., 2022). This is a serious issue, as depression can significantly disrupt daily life, leading to poor academic performance, insomnia, suicidal tendencies, and even school dropouts. Schools play a crucial role in students' mental health development. According to Liu & Wang (2024), the high prevalence of depression and anxiety has led to a noticeable decline in student satisfaction with university life. Addressing students' needs and fostering a supportive environment are essential for both their well-being and the success of academic institutions. Understanding the socio-economic, lifestyle, and health-related factors contributing to student depression is vital. In this paper, we analyze a student depression dataset to examine how academic pressure, financial stress, study duration, student satisfaction and family history of mental illness can increase the risk of depression among students.

1.1 Dataset cleaning and Preprocessing

The dataset for this project, obtained from Kaggle [3], contains comprehensive information on individuals' personal and lifestyle factors. It is structured to facilitate analysis in health, lifestyle, and socio-economic contexts. The author of the original dataset is Shodolamu Opeyemi [4]. The table 1 shows the dataset columns and its respective description:

Table 1: Dataset Structure

Column Name	***Description***
id	*ID of the student.*
Gender	*Gender of the student (male, female).*
Age	*Age of the student in years.*
City	*City where student resides.*
Profession	*Individual Profession.*
Academic Pressure	*How pressure does the student feel rewarding studies?*
Work Pressure	*How pressure does the student feel rewarding work?*
CGPA	*Student CGPA (grades).*
Study Satisfaction	*How happy is the student with studies?*
Job Satisfaction	*How happy is the student with work?*
Sleep Duration	*Quality of sleep (e.g., good, fair, poor).*
Dietary Habits	*Dietary habits (e.g., healthy, moderate, unhealthy).*
Degree	*Degree that the students is working on.*
Suicidal thoughts?	*Whether the student has suicidal thoughts. (yes/no).*
Work/Study Hours	*Number of Hours dedicated to study.*
Financial Stress	*Whether the student has Financial Stress. (yes/no).*
Family History of Mental Illness	*Does student have family with Mental Health Illness (yes/no)*
Depression	*Does student have depression (yes/no)*

Table 1 indicates that the original dataset structure consists of 18 columns. However, for the specific analysis conducted in this project, certain columns are not relevant and can be removed. The first step in data cleaning involves eliminating unnecessary columns that do not contribute to the objectives of the analysis. Specifically, columns related to employment, such as **Profession, Work Pressure, and Job Satisfaction**, are excluded as they are not central to the research focus. Additionally, the **ID** and **City** columns are not essential for this study and have been removed. Figure 1 presents the refined dataset, containing only the variables necessary for the analysis.

Gender	Age	City	Academic Pressure	CGPA	Study Satisfaction	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/Study Hours	Financial Stress	Family History of Mental Illness	Depression
Male	33.0	Visakhapatnam	5.0	8.97	2.0	5-6 hours	Healthy	B.Pharm	Yes	3.0	1.0	No	1
Female	24.0	Bangalore	2.0	5.90	5.0	5-6 hours	Moderate	BSc	No	3.0	2.0	Yes	0
Male	31.0	Srinagar	3.0	7.03	5.0	Less than 5 hours	Healthy	BA	No	9.0	1.0	Yes	0
Female	28.0	Varanasi	3.0	5.59	2.0	7-8 hours	Moderate	BCA	Yes	4.0	5.0	Yes	1

Figure 1: Dataset Overview

The next step in the data cleaning process involves renaming certain columns to enhance clarity and ensure a more intuitive interpretation of the dataset. Additionally, CGPA values have been adjusted to align with the Canadian grading scale.

Gender	Age	Academic Pressure	CGPA	Study Satisfaction	Sleep Duration	Dietary Habits	Suicidal_Thoughts	Study_Hours	Financial Stress	Fam_Mental_Hist	Depression
Male	33	5	4.485	2	5-6 hours	Healthy	Yes	3	1.0	No	1
Female	24	2	2.950	5	5-6 hours	Moderate	No	3	2.0	Yes	0
Male	31	3	3.515	5	Less than 5 hours	Healthy	No	9	1.0	Yes	0
Female	28	3	2.795	2	7-8 hours	Moderate	Yes	4	5.0	Yes	1
Female	25	4	4.065	3	5-6 hours	Moderate	Yes	1	1.0	No	0

Figure 2 Cleaned dataset

Figure 2 presents the cleaned dataset. The next step is to verify the type of data that the columns hold and converted to the appropriate one. For instance, the age columns need to be converted to a group to have a better structure for further analysis. Table 2 shows the name of the groups and the age range that the groups hold. The name of the columns that holds these values is called 'Age_Group'.

Table 2 Age Column to Age_Group

Group Name for AGE categories	***Range***
Group_1	*Less that 28 years*
Group_2	*From 29 yeas to 38 years*
Group_3	*From 29 yeas to 48 years*
Group_4	*From 29 yeas to 59 years*

In addition to that, the existing columns with groups categories such as 'Sleep Duration' and 'Dietary Habits' contained groups called 'others' that are not relevant to the analysis therefore the fields with these values were dropped leaving only those values that are relevant for the analysis. The same situation happened for the column named 'Academic Pressure', containing 1 value of 0, that value was dropped due to its irrelevance for the analysis. The final step is the verification of the final dataset and its respective structure.

```
<class 'pandas.core.frame.DataFrame'>
Index: 27856 entries, 0 to 27900
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Gender               27856 non-null  object
 1   Academic Pressure    27856 non-null  int64
 2   CGPA                 27856 non-null  float64
 3   Study Satisfaction   27856 non-null  int64
 4   Sleep Duration       27856 non-null  object
 5   Dietary Habits       27856 non-null  object
 6   Suicidal_Thoughts    27856 non-null  object
 7   Study_Hours          27856 non-null  int64
 8   Financial Stress     27856 non-null  float64
 9   Fam_Mental_Hist      27856 non-null  object
 10  Depression           27856 non-null  int64
 11  Age_Group            27856 non-null  category
dtypes: category(1), float64(2), int64(4), object(5)
memory usage: 2.6+ MB
```

Figure 3 Dataset Information

Figure 3 shows the information of the dataset listing all the columns. The final dataset consists of **27,856 rows**, where each row represents an independent student. It contains **12 columns**, each with distinct attributes and data types. Specifically, the dataset includes:

- **2 columns** with a **float** data type,
- **4 columns** with an **integer** data type,
- **5 columns** with an **object (string)** data type, and
- **1 column** with a **category** data type.

Moreover, the cleaned dataset contains **duplicates** and **missing values**. The table below compares the structure of the original dataset with the cleaned dataset, highlighting the modifications made during the preprocessing steps.

Table 3 Data Cleaning Summary

Original Dataset	***Cleaned Dataset***
18 Columns	12 Columns
27901 Rows	27856 Rows
Age in INT	Age in category (Age_Group)
Dietary Habits include 'Other' category	Dietary Habits Removed 'Other' category
Sleep Duration include 'Other' category	Sleep Duration Removed 'Other' category

1.2 Research Questions:

This analysis aims to explore the key factors influencing student depression, particularly focusing on academic pressure, lifestyle habits, financial stress, and family history of mental illness. By analyzing these factors, the research seeks to uncover patterns that contribute to depression among students. Additionally, the study investigates the potential for predictive modeling to identify students at higher risk of developing depression. The following research questions guide the analysis:

1. **Academic Pressure & Mental Health**
    - How does academic pressure affect student mental health?
2. **Lifestyle Factors & Depression**
    - Is there a relationship between sleep duration and depression?
3. **Stress & Mental Well-being**
    - How strongly does financial stress impact depression in students?
4. **Family & Mental Health History**
    - Does having a family history of mental illness increase the risk of depression?
5. **Predictive Modeling & Insights**
    - How can we predict student depression using machine learning?

2. Exploratory Data Analysis (EDA)

The exploratory data analysis (EDA) aims to examine the underlying structure of the dataset and identify potential patterns or trends. This process provides valuable insights into the relationships between variables and helps uncover factors contributing to specific outcomes. Recognizing these patterns can enhance our understanding of the events occurring within the dataset. Even if direct causal relationships are not established, EDA can offer important insights that may help mitigate risk factors and inform data-driven decision-making.

This analysis focuses on key variables related to **socio-economic status, lifestyle habits, and mental health**, exploring their interactions and potential impact on student well-being.

2.1 Depression cases

The initial step in this analysis involves examining the confirmed cases of depression in the dataset. This foundational understanding is crucial for assessing the prevalence of depression among students and identifying key contributing factors. To facilitate this exploration, **R and Python** were utilized as analytical tools, ensuring an efficient and systematic approach to data exploration. **Figure 4** shows the total number of depression cases within the dataset.

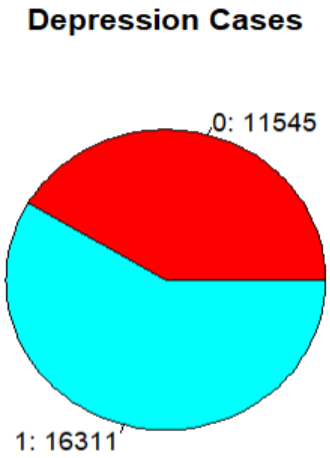

Figure 4 Total of Depression Cases

In **Figure 4**, the value **0** represents students who have **not** been diagnosed with depression, while the value **1** indicates students with a **confirmed depression diagnosis**. The dataset comprises **11,545 students without depression** and **16,311 students diagnosed with depression**, indicating that **over 50% of students in this dataset experience depression**.

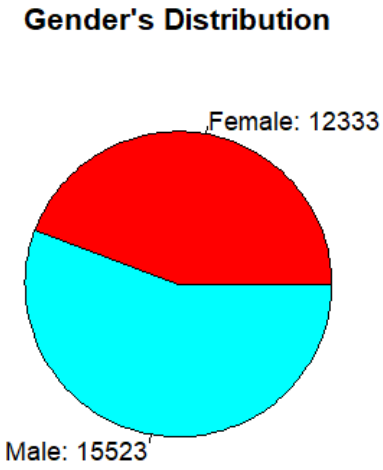

Figure 5 Male and Female Count

In addition to examining overall depression prevalence, we analyzed how depression is distributed among male and female students. **Figure 5** presents the actual distribution of depression cases across these demographics, providing insights into potential differences in mental health outcomes between male and female students. The dataset includes **12,333 female students** and **15,523 male students**, indicating a higher representation of males. It is important to note this gender imbalance when interpreting the results.

To understand the relevance of the distribution the next step is to display the distribution of positive cases between males and females, if the number of males in the study is bigger, then an easy assumption will be that the positive cases will be bigger amongst the males as well. This is just an assumption following the original numbers and distributions. The Figure 6 shows the real distribution among Male and females.

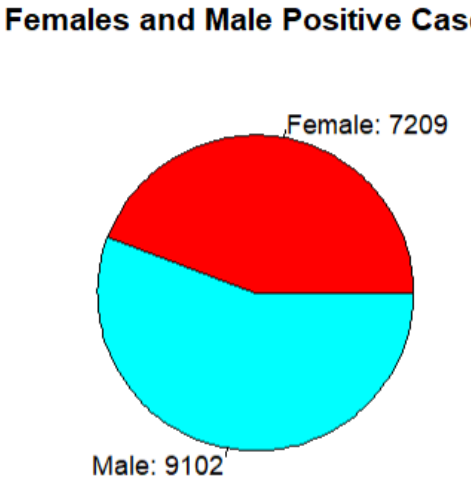

Figure 6 Positive Cases Male Vs Female

From **Figure 6** the assumption stipulated before has been confirm. More than 50% of positive depression cases in students are males. Although common knowledge suggests that females, being more emotional, are more vulnerable to depression, the data indicates otherwise. The ability to express their emotions more freely might be a form of relief for females. On the other hand, males, often forced to control and suppress their emotions due to social prejudices, might be more adversely affected. Another interesting factor is that the distribution between the total females and males in the study follows similar structure when regarding the distribution in only positive cases having approximately 56% males and 44% females in the positive cases.

These insights highlight the importance of considering **social and cultural factors** when analyzing mental health trends, as gender norms may play a significant role in how depression manifests and is reported among students. The next phase of the analysis focuses on identifying key factors associated with depression. These factors are categorized into four primary groups:

- Academic Pressure & mental Health
- Lifestyle Factors
- Stress & Mental Well-being
- Family & Mental Health History.

Groups	***Factors / Columns***
Lifestyle Factors	*Dietary Habits, Sleep Duration, Age, Gender*
Stress & Mental Well-being	*Financial Stress, Work/Study Hours, Academic Pressure*
Family & Mental Health History.	*Suicidal_Thoughts, Fam_Mental_Hist*

2.2 Lifestyle Factors

In this section the analysis will be based in 4 different groups to understand how the factor influence depression in students. The analysis has been done using charts.

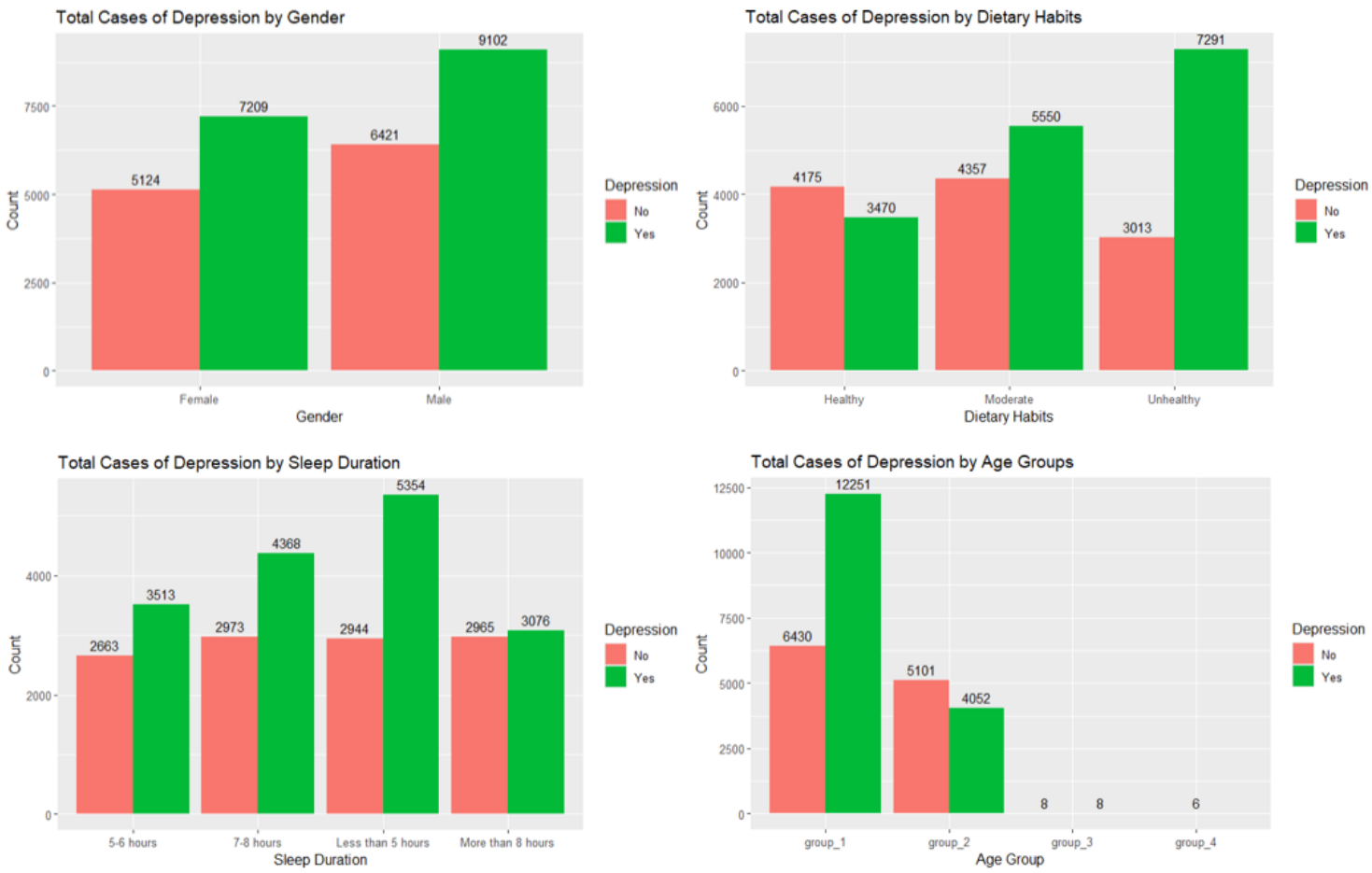

Figure 7 Lifestyle Factors vs Depression

Figure 7 reveals several important patterns related to **lifestyle factors and depression** among students:

1. **Diet and Depression:**
    - Students with **unhealthy dietary habits** are more likely to be diagnosed with depression. This aligns with research suggesting that poor nutrition can negatively impact mental health.
2. **Sleep Duration and Depression:**
    - Students who **sleep less than five hours per night** have a higher likelihood of experiencing depression. Sleep deprivation has been widely linked to increased stress, anxiety, and mood disorders.
3. **Age and Depression Risk:**

- Most students in the dataset are **under the age of 38**, and within this age group, depression is more prevalent. This may indicate that younger students, who often face significant academic and career pressures, are at greater risk.

4. **Gender and Depression:**
    - Since the dataset contains **a higher proportion of male students**, the findings suggest that males are more likely to be diagnosed with depression. This supports earlier observations regarding gender differences in emotional expression and coping mechanisms.

These insights highlight the importance of **lifestyle choices, sleep habits, age, and gender dynamics** in understanding student mental health. Further analysis can help determine the extent to which these factors contribute to depression and how interventions can be tailored to mitigate risks

2.3 Stress & Mental Well-being Factors

In this section the analysis will be based on 4 different groups to understand how the factor influence depression in students. The analysis has been done using charts.

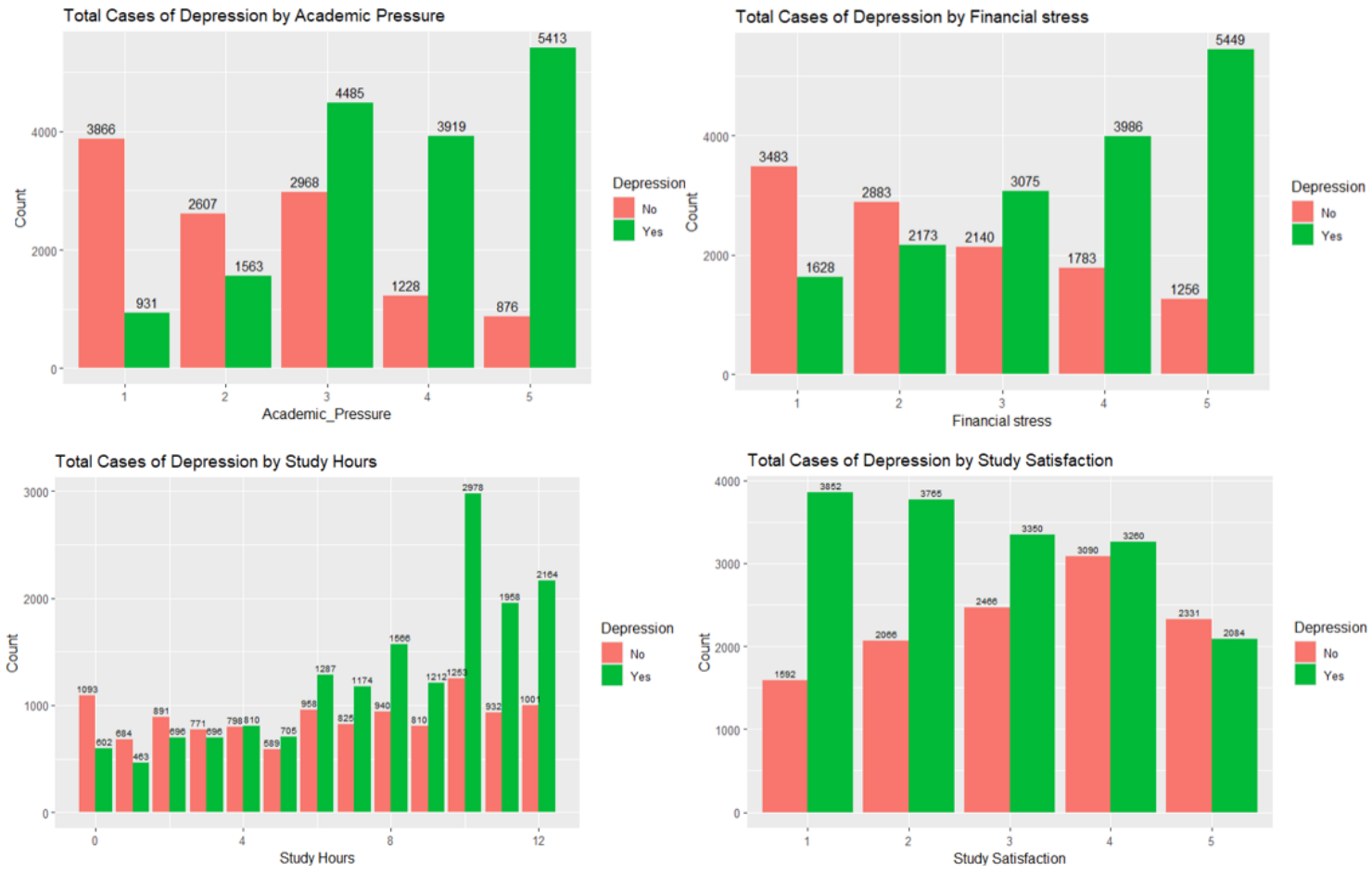

Figure 8 Stress Factors vs Depression

Figure 8 provides important observations on how **stress factors influence depression among students**:

1. **Academic Pressure and Depression:**
    - Students who experience **higher levels of academic pressure** are more likely to be diagnosed with depression. This suggests that excessive academic demands can negatively impact mental health.
2. **Financial Stress and Depression:**
    - Students who **face financial stress** have a higher likelihood of depression. Financial instability can contribute to anxiety, uncertainty, and increased psychological distress.
3. **Study Hours and Depression:**

- Students who dedicate **longer hours to studying** are more prone to depression. However, an interesting trend emerges **after 10 hours of study per day, depression diagnoses decrease**. This could be due to a smaller number of students studying for extended hours, or it may indicate that extremely dedicated students have better coping mechanisms.

4. **Study Satisfaction and Depression:**
    - Students with **low study satisfaction** are more likely to be diagnosed with depression. A negative academic experience, including dissatisfaction with coursework or career prospects, may contribute to mental health struggles.

These findings highlight the **critical role of stress both academic and financial in shaping student mental health**. Further investigation is needed to explore how institutional and personal coping strategies can help mitigate these risks.

2.4 Mental Health-Factors

In this section the analysis will be based in 2 different groups to understand how the factor influence depression in students. The analysis has been done using charts.

Figure 9 Metal Health vs Depression

This section examines how **mental health-related factors** influence depression among students. The analysis is divided into two groups to provide a clearer understanding of their impact, using visual representations such as charts to highlight key trends.

1. **Suicidal Thoughts and Depression:**
    - Students who **report having suicidal thoughts** are significantly more likely to be diagnosed with depression. This finding reinforces the strong link between depression and suicidal ideation, emphasizing the need for early mental health interventions.
2. **Family History of Mental Health Disorders and Depression:**
    - Students with **a family history of mental illness** are slightly more likely to be diagnosed with depression. This suggests a potential genetic or environmental influence, where individuals with close relatives experiencing mental health disorders may be at a higher risk themselves.

These findings underscore the importance of **early screening, mental health support, and targeted interventions** for students facing psychological distress or familial predispositions to mental health challenges.

2.5 Preliminary Insights from Exploratory Data Analysis (EDA)

While it is essential to conduct a deeper statistical analysis to establish relationships between **independent variables** and the **response variable (depression diagnosis)**, the exploratory data analysis (EDA) provides preliminary observations. Based on the trends identified, we can make the following statements:

- **Students who experience suicidal thoughts** are significantly more likely to be diagnosed with depression.
- **Students in Group 1 (ages below 28)** are the most affected by depression, indicating a higher prevalence in younger individuals.
- **Students facing financial stress** are at a greater risk of being diagnosed with depression, suggesting a strong link between economic hardship and mental well-being.
- **Students experiencing high academic pressure** are more likely to be diagnosed with depression, reinforcing the impact of academic stress on mental health.

Although these insights offer valuable preliminary findings, **further statistical analysis is necessary** to confirm or reject these assumptions. More advanced methods, such as regression analysis or machine learning models, can provide a deeper understanding of the relationships between these factors and depression diagnoses.

3. Statistical Analysis

In this section, the dataset has been structured and prepared for model selection. Key data elements were previously analyzed and are summarized below as a reference:

- **Total students:** 27,856
- **Total positive depression cases:** 16,311
- **Total negative depression cases:** 11,545
- **Total male students:** 15,523
    - Positive cases (male): 9,102
    - Negative cases (male): 6,421
- **Total female students:** 12,333
    - **Positive cases (female):** 7,209
    - **Negative cases (female):** 5,124

Given this distribution, the dataset must be **properly divided** to ensure a balanced and representative sample for model training. To achieve this, a **stratified random sampling** approach will be used. This method ensures that the proportion of **male and female students, as well as positive and negative depression cases, are maintained within the sample**.

The next step in this process involves defining the **proportions for each stratum**, ensuring that the selected sample accurately reflects the structure of the full dataset. This will help maintain the integrity of the analysis and improve the reliability of predictive models.

The Figure 10 shows and overall idea of the distribution and the stratum proportion that each combination of gender and depression status have over the study.

Stratum Proportion:

- Positive Male = 0.3268
- negative Male =0.2305
- positive Female =0.2588
- negative Female =0.1840

```r
# Get the Total Count of Students in the Study
total_Cases <- nrow(depression_dataset)

# Separate the Study in Positive and Negative Depression Cases
negative_Cases <- depression_dataset[depression_dataset$Depression == 'No', ]
positive_Cases <- depression_dataset[depression_dataset$Depression == 'Yes', ]

# Get the Stratum Proportion per Gender and Positive or Negative Cases
#Negative
male_negative <- nrow(negative_Cases[negative_Cases$Gender=='Male',])/total_Cases
female_negative <-nrow(negative_Cases[negative_Cases$Gender=='Female',])/total_Cases

#Positive
male_positive <- nrow(positive_Cases[positive_Cases$Gender=='Male',])/total_Cases
female_positive <-nrow(positive_Cases[positive_Cases$Gender=='Female',])/total_Cases

male_positive
male_negative
female_positive
female_negative
```

```
[1] 0.3267519
[1] 0.2305069
[1] 0.2587952
[1] 0.183946
```

Figure 10: Stratum Proportions

3.1 Stratified Random Sampling

Stratified simple sampling is a method of sampling that involves dividing a population into smaller groups, known as strata, that share similar characteristics. Within each stratum, a simple random sample is taken. This ensures that each subgroup is adequately represented in the overall sample, leading to more accurate and reliable results. It is particularly useful when there are distinct subgroups within a population that may have different behaviors or attributes. The main benefit of stratified sampling is that it increases the precision of the overall estimates by reducing sampling variability. This method also allows for more detailed subgroup analysis, providing insights into the characteristics and trends within each stratum.

The next step is to generate the stratified random sample.

```r
# Create a stratified sample based on Gender and Depression
set.seed(2025)  # For reproducibility
strata_columns <- c("Gender", "Depression")
stratified_sample <- depression_dataset %>%
  group_by_at(strata_columns) %>%
  sample_frac(size = 0.50)

# Check the stratified sample
print(table(stratified_sample$Gender, stratified_sample$Depression))
```

```
          No  Yes
Female 2562 3604
Male   3210 4551
```

Figure 11 Stratified Random Sample

A few clarifications about the code snippet above in figure 11. The stratification is applied to the columns Depression and Gender. This ensures that the sample is filtered through these columns, resulting in values that are a combination of these attributes, as previously shown. The sample size is half of the original dataset size. The decision to use 50% of the original size aims to have a sufficiently large dataset for further division between the training and testing sets for analysis. This way, we can still have a substantial amount of data to analyze, but it remains small enough to compute efficiently. Also, the code uses the "set. Seed" command to set the create reproducibility allowing same results in different devices. The next step is to proceed to generate the two used dataset for training and for testing.

3.2 Split Data into Training and Testing

Splitting data is the action of dividing the working data in two parts:

- Training set (75%)
- Test set (25%)

Splitting data is a common practice in machine learning to evaluate the performance of a model. The training set is used to train the model, allowing it to learn patterns and relationships within the data. The test set, which is kept separate and not seen by the model during training, is used to assess the model's performance. This helps ensure that the model can generalize well to new, unseen data, and not just memorize the training data. The main benefit of this approach is that it provides a reliable estimate of the model's accuracy and robustness, helping to prevent overfitting and underfitting. The standard data split is 75% for training and 25% for testing.

```r
# Split the stratified sample into training and testing sets (75% training, 25% testing)
trainIndex <- createDataPartition(stratified_sample$Depression, p = 0.75, list = FALSE)
train_data <- stratified_sample[trainIndex,]
test_data <- stratified_sample[-trainIndex,]

original_data_set_percentage75 <- nrow(stratified_sample)*0.75
original_data_set_percentage25 <- nrow(stratified_sample)*0.25

  # Check the splits
# 75% of the stratified sample
print (original_data_set_percentage75)
print(nrow(train_data))

# 25% of the stratified sample
print (original_data_set_percentage25)
print(nrow(test_data))
```

```
[1] 10445.25
[1] 10446
[1] 3481.75
[1] 3481
```

Figure 12 Split Stratified Sample

From figure 12 the new datasets had been created, having a training and a testing sets to conduct further analysis and prove the model's behaviour and accuracy. The distribution for the training and testing data sets are 75% and 25% respectively form the stratified sample which is 50% of the original dataset. The printed values in the code compare the total number of rows in the stratified sample with the counts of rows in the training and testing splits. This step was performed to ensure that the row counts in the datasets align correctly.

3.3 Logistic Regression Model

When using a logistic model, it is essential to consider a few key elements or conditions. The logistic model is appropriate when the response variable is binary, meaning it takes on values between 0 and 1. This is precisely what is needed for the analysis, as the response variable is 'Depression,' recorded as '0' and '1,' 'Yes' and 'No,' or 'True' and 'False.' Therefore, treating the response variable as binary makes the logistic model a suitable choice for this analysis.

Another important factor to consider is how we evaluate the response variable. Although the values between 0 and 1 are continuous, for the sake of this analysis all values greater than 0.5 were assigned as '1' and all values less than or equal to 0.5 as '0.' This binary classification allows to use the logistic model effectively to analyze the data.

One final element to consider is the appropriate classification of the variables (columns). It is crucial to treat these variables as intended. Specifically, columns that hold numeric values but represent categorical options should be converted to factors. Even though they contain numeric values, these numbers are merely grouped options, and

thus, they need to be treated as factors for the model to accurately interpret and categorize them. The below are the steps to implement Logistic Regression:

1. **Verify Columns Classification:**

```r
```{r}
str(train_data)
```
```

```
gropd_df [10,446 × 12] (S3: grouped_df/tbl_df/tbl/data.frame)
 $ Gender            : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
 $ Academic_Pressure : Factor w/ 5 levels "1","2","3","4",..: 1 1 3 1 2 2 1 2 3 1 ...
 $ CGPA              : num [1:10446] 2.93 4.95 4.02 4.61 3.69 ...
 $ Study_Satisfaction: Factor w/ 5 levels "1","2","3","4",..: 2 5 5 3 4 1 1 5 3 3 ...
 $ Sleep_Duration    : Factor w/ 4 levels "5-6 hours","7-8 hours",..: 3 4 1 1 4 1 3 1 2 2 ...
 $ Dietary_Habits    : Factor w/ 3 levels "Healthy","Moderate",..: 1 3 1 2 1 1 2 1 1 2 ...
 $ Suicidal_Thoughts : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 2 1 2 1 ...
 $ Study_Hours       : int [1:10446] 7 3 0 0 1 12 5 0 9 11 ...
 $ Financial_stress  : Factor w/ 5 levels "1","2","3","4",..: 2 3 3 1 5 1 3 1 4 4 ...
 $ Fam_Mental_Hist   : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 2 1 2 ...
 $ Depression        : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Age_Group         : Factor w/ 4 levels "group_1","group_2",..: 1 2 1 2 1 1 1 1 1 1 ...
```

Figure 13 Columns Classification

From figure 13, it is possible to verify the status of each column and how will they be treated when entered in the model command. From all the columns only two columns are numeric, and the rest are categorical columns.

2. **Fit the model:**

The Figure 14 shows a few important facts about the model. The CGPA and Gender predictors are insignificant when the alpha level is set at 0.05. The model verifies several assumptions about the behavior of the data. For instance, when the financial stress factor increases to a higher level, the coefficients for that predictor also increase. This leads to the conclusion that individuals experiencing higher financial stress are more likely to be diagnosed with depression. A similar pattern is observed with the academic pressure predictor; as academic

pressure levels increase, so do the coefficients. Conversely, when the level of student satisfaction increases, the coefficients decrease, which in turn reduces the likelihood of being diagnosed with depression.

3. **Verify multicollinearity:**

```
```{r}
# Fit the logistic regression model
logistic_Model <- glm(Depression ~.,data = train_data, family = binomial)

# Summarize the model
summary(model)
```
```

```
Call:
glm(formula = Depression ~ ., family = binomial, data = train_data)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -5.077482   0.233021 -21.790  < 2e-16 ***
GenderMale                      -0.026681   0.060324  -0.442 0.658280
Academic_Pressure2               1.068665   0.102801  10.395  < 2e-16 ***
Academic_Pressure3               1.897728   0.092522  20.511  < 2e-16 ***
Academic_Pressure4               2.712256   0.104268  26.012  < 2e-16 ***
Academic_Pressure5               3.478564   0.110740  31.412  < 2e-16 ***
CGPA                             0.063501   0.040744   1.559 0.119108
Study_Satisfaction2             -0.459661   0.098593  -4.662 3.13e-06 ***
Study_Satisfaction3             -0.497498   0.097122  -5.122 3.02e-07 ***
Study_Satisfaction4             -0.805318   0.095169  -8.462  < 2e-16 ***
Study_Satisfaction5             -1.072992   0.102911 -10.426  < 2e-16 ***
Sleep_Duration7-8 hours          0.063094   0.085391   0.739 0.459974
Sleep_DurationLess than 5 hours  0.292438   0.084317   3.468 0.000524 ***
Sleep_DurationMore than 8 hours -0.355809   0.089151  -3.991 6.58e-05 ***
Dietary_HabitsModerate           0.473761   0.073275   6.466 1.01e-10 ***
Dietary_HabitsUnhealthy          1.180351   0.076332  15.463  < 2e-16 ***
Suicidal_ThoughtsYes             2.496880   0.064177  38.906  < 2e-16 ***
Study_Hours                      0.122943   0.008179  15.031  < 2e-16 ***
Financial_stress2                0.451806   0.095585   4.727 2.28e-06 ***
Financial_stress3                1.026654   0.093945  10.928  < 2e-16 ***
Financial_stress4                1.486964   0.095078  15.639  < 2e-16 ***
Financial_stress5                2.242554   0.098192  22.839  < 2e-16 ***
Fam_Mental_HistYes               0.314024   0.059913   5.241 1.59e-07 ***
Age_Groupgroup_2                -0.962288   0.063613 -15.127  < 2e-16 ***
Age_Groupgroup_3                -0.556789   1.570074  -0.355 0.722870
Age_Groupgroup_4               -11.469077 139.130423  -0.082 0.934302
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14173.7  on 10445  degrees of freedom
Residual deviance:  7245.8  on 10420  degrees of freedom
AIC: 7297.8

Number of Fisher Scoring iterations: 10
```

Figure 14 Fit Logistic Model/ Summary

Multicollinearity occurs when predictors in a model are highly correlated with each other, which can lead to instability in the coefficient estimates and make it difficult to determine the individual effect of each predictor. That is why is important to keep record of it. The list below shows a brief interpretation of the relevance on multicollinearity.

- **VIF = 1**: No correlation between the predictor and other predictors.
- **1 < VIF < 5**: Moderate correlation, typically not a cause for concern.
- **VIF ≥ 5**: High correlation, indicating significant multicollinearity.
- **VIF > 10**: Very high correlation, often considered a serious problem that requires correction.

```r
library(car)
vif(logistic_Model)
```

	GVIF	Df	GVIF^(1/(2*Df))
Gender	1.011975	1	1.005970
Academic_Pressure	1.125542	4	1.014893
CGPA	1.011823	1	1.005894
Study_Satisfaction	1.048684	4	1.005960
Sleep_Duration	1.022292	3	1.003681
Dietary_Habits	1.041278	2	1.010163
Suicidal_Thoughts	1.082380	1	1.040375
Study_Hours	1.024995	1	1.012421
Financial_stress	1.069461	4	1.008430
Fam_Mental_Hist	1.006080	1	1.003035
Age_Group	1.029187	3	1.004806

Figure 15 Multicollinearity Test

Figure 15 shows the result of the multicollinearity analysis and from the picture is evident that the values of the predictors range between 1 and 4 staying lower than 5. Therefore, it is possible to conclude that the predictors are not high correlated, and no further process is needed for this section.

From the modeling fitting we notice that there are two predictors that are insignificant for the model. Therefore, we could potentially have two logistic models.

- Original Model: that includes all the predictors
- Reduce Model: that holds only the predictors that are significant

The correspondent process for the original model had been done. The reduced model needs to also be loaded and verify its multicollinearity.

1. **Load Reduce Logistic Model**

```r
# Fit the logistic regression model
reduce_Logistic_Model <- glm(Depression ~ Academic_Pressure +  Study_Satisfaction +
                               Sleep_Duration + Dietary_Habits + Suicidal_Thoughts +
                             Study_Hours + Financial_stress + Fam_Mental_Hist +
                               Age_Group, data = train_data, family = binomial)

# Summarize the model
summary(reduce_Logistic_Model)
```

Figure 16 Reduce Logistic Model

Figure 16 shows the reduced logistic model with only the predictors that were significant for the model. Figure 17 shows the summary of the Reduced Logistic Model. From figure 17, It is possible to verify that the new predictors are all significant for the model.

```
Call:
glm(formula = Depression ~ Academic_Pressure + Study_Satisfaction +
    Sleep_Duration + Dietary_Habits + Suicidal_Thoughts + Study_Hours +
    Financial_stress + Fam_Mental_Hist + Age_Group, family = binomial,
    data = train_data)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -4.842664   0.166377 -29.107  < 2e-16 ***
Academic_Pressure2               1.067847   0.102737  10.394  < 2e-16 ***
Academic_Pressure3               1.895354   0.092459  20.499  < 2e-16 ***
Academic_Pressure4               2.707400   0.104132  26.000  < 2e-16 ***
Academic_Pressure5               3.473145   0.110597  31.404  < 2e-16 ***
Study_Satisfaction2             -0.460333   0.098574  -4.670 3.01e-06 ***
Study_Satisfaction3             -0.502662   0.097045  -5.180 2.22e-07 ***
Study_Satisfaction4             -0.807683   0.095137  -8.490  < 2e-16 ***
Study_Satisfaction5             -1.081502   0.102684 -10.532  < 2e-16 ***
Sleep_Duration7-8 hours          0.064075   0.085370   0.751 0.452924
Sleep_DurationLess than 5 hours  0.292580   0.084276   3.472 0.000517 ***
Sleep_DurationMore than 8 hours -0.359162   0.089101  -4.031 5.55e-05 ***
Dietary_HabitsModerate           0.473115   0.073258   6.458 1.06e-10 ***
Dietary_HabitsUnhealthy          1.178791   0.076143  15.481  < 2e-16 ***
Suicidal_ThoughtsYes             2.497810   0.064161  38.930  < 2e-16 ***
Study_Hours                      0.122726   0.008172  15.018  < 2e-16 ***
Financial_stress2                0.450775   0.095564   4.717 2.39e-06 ***
Financial_stress3                1.028015   0.093934  10.944  < 2e-16 ***
Financial_stress4                1.490379   0.095031  15.683  < 2e-16 ***
Financial_stress5                2.242510   0.098150  22.848  < 2e-16 ***
Fam_Mental_HistYes               0.314267   0.059903   5.246 1.55e-07 ***
Age_Groupgroup_2                -0.962550   0.063602 -15.134  < 2e-16 ***
Age_Groupgroup_3                -0.615111   1.576099  -0.390 0.696333
Age_Groupgroup_4               -11.434775 139.146444  -0.082 0.934505
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 14173.7  on 10445  degrees of freedom
Residual deviance:  7248.4  on 10422  degrees of freedom
AIC: 7296.4

Number of Fisher Scoring iterations: 10
```

Figure 17 Summary Reduce Logistic Model

2. **Verify multicollinearity:** From figure 18, we confirm that the predictors are not high correlated.

```
```{r}
vif(reduce_Logistic_Model)
```
```

```
                     GVIF Df GVIF^(1/(2*Df))
Academic_Pressure  1.121936  4        1.014486
Study_Satisfaction 1.042953  4        1.005271
Sleep_Duration     1.020801  3        1.003437
Dietary_Habits     1.034161  2        1.008433
Suicidal_Thoughts  1.082465  1        1.040416
Study_Hours        1.023700  1        1.011781
Financial_stress   1.067063  4        1.008147
Fam_Mental_Hist    1.006141  1        1.003066
Age_Group          1.028476  3        1.004691
```

Figure 18 Reduce Logistic Model multicollinearity

3.3.1 Logistic Regression Model Testing

There are two available options: the original model, which includes the insignificant predictors Gender and CGPA, and a reduced model where these insignificant predictors had been removed. For the sake of the analysis, the next step involves applying both models and comparing their respective behaviors to highlight any interesting results.

Full Logistic model:

Using the created full logistic model, the test dataset previously created its going to be used for this test.
From the results shown in figure 19, there are 2 main results:

````
**Applying Logistic Model to Test Dataset**
```{r}
predictions <- predict(logistic_Model, newdata=test_data, type='response')
pred_class <- ifelse(predictions >= 0.5, 1, 0)
```

**Create a confusion matrix**
```{r}
conf_matrix <- table(Predicted = pred_class, Actual = test_data$Depression)
print(conf_matrix)
```
````

```
          Actual
Predicted   No  Yes
        0 1143  243
        1  300 1795
```

````
**Percentage of correct predictions**
```{r}
correct_predictions <- sum(diag(conf_matrix)) / sum(conf_matrix)
print(paste("Overall Correct Predictions Percentages: ", correct_predictions))
```
````

```
[1] "Overall Correct Predictions Percentages:  0.844010341855789"
```

Figure 19: Full Logistic Model Test

Out of the 3481 rows in the test dataset the model successfully predicted 2931 depression cases. Therefore 550 depression cases are false positive and false negative. The accuracy of this model is about 84.4%. In other words, 84.4% of the times, it predicts correctly.

Reduce Logistic Model:

Following the same steps from the previous testing but in this case using the Reduce Logistic Model.

The next step is testing the reduce model with test dataset.

```r
predictions_reduce <- predict(reduce_Logistic_Model, newdata=test_data, type='response')
pred_class_reduce <- ifelse(predictions_reduce >= 0.5, 1, 0)
```

Create a confusion matrix

```r
Predicted = pred_class_reduce
Actual = test_data$Depression
conf_matrix_reduce <- table(Predicted,Actual )
print(conf_matrix_reduce)
```

```
         Actual
Predicted   No  Yes
        0 1144  248
        1  299 1790
```

Percentage of correct predictions

```r
correct_predictions_reduce <- sum(diag(conf_matrix_reduce)) / sum(conf_matrix_reduce)
print(paste("Overall Fraction of Correct Predictions: ", correct_predictions_reduce))
```

```
[1] "Overall Fraction of Correct Predictions:  0.84286124676817"
```

Figure 20 Reduced Logistic Model Test

From the results shown in figure 20, we conclude that.

- Out of the 3481 rows in the test dataset the model successfully predicted 2934 depression cases. Therefore 547 depression cases are false positive and false negative.
- The accuracy of this model is about 84.3%. In other words, 84.3% of the times, it predicts correctly.

Accuracy for both Logistic and Reduced Logistic Models:

After running the same test in both logistic models, it is easier to compare them and display an objective result:

```r
# Create a data frame to store model performance
model_results <- data.frame(
  Model = c("Logistic Regression", "Reduce Logistic Model"),
  Correct_rate_Predictors = c(
    correct_predictions,
    correct_predictions_reduce
  )
)

# Display the results
print(model_results)
```

Description: df [2 × 2]

Model <chr>	Correct_rate_Predictors <dbl>
Logistic Regression	0.8440103
Reduce Logistic Model	0.8428612

Figure 21: Comparison of Logistic Models

Figure 21 shows the comparison between the two logistic models. It is evident that the original model behaves better than the reduced model when evaluated or being used for prediction purposes. However, the reduce model is less complex having two less predictors. According to the needs of the user and the actual application of the model the decision of which model is better depended on the mentioned factor.

3.4 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a supervised machine learning technique used for classification and dimensionality reduction. It identifies a linear combination of features that optimally separates multiple classes by maximizing the ratio of between-class variance to within-class variance. LDA constructs the within-class scatter matrix, representing variations within each class, and the between-class scatter matrix, capturing differences between class means, to determine the most effective projection direction. Unlike Principal Component Analysis (PCA), which prioritizes variance, LDA preserves information crucial for distinguishing classes. It is widely applied in pattern recognition, medical diagnosis, and face recognition, though its effectiveness depends on the assumption that data is normally distributed with equal covariance across classes, which should be validated before use.

In this section we employ the LDA model to the stratified sample created in the statistical analysis section as shown in figure 12 to compare its performance with other models. Unlike logistic regression, which is mainly for binary classification, LDA works for both binary and multiple classes. The dependent variable here is depression which we have categorised into two levels, “yes” and “no”.

LDA model and QDA model needs to meet some data assumptions those are:

- Normality
- Linear independency

To verify these assumptions, we can apply some verification methods as previously done for multicollinearity for the logistic model.

1. **Linear Independence:**

From previous analysis, the confirmation of linear independency was confirmed. Please check Figure 15 for confirmation.

2. **Normality:**

To verify normality, we implemented two normality tests QQ-plot and Shapiro test.

3. **QQ Plot:**

A Quantile-Quantile (QQ) plot is a graphical tool to assess if a dataset follows a specified distribution (commonly the normal distribution). It compares the quantiles of the dataset against the quantiles of the theoretical distribution. If the data is normally distributed, the points on the QQ plot will approximately lie along a straight line. Deviations from this line indicate departures from normality. Figure 22 shows the behaviour of the data in the dataset does not follow a normal distribution.

```
library(ggpubr)
# Q-Q plot to check normality of CGPA for each Depression class
ggqqplot(train_data, x = "Study_Hours", facet.by = "Depression")
```

```
# Q-Q plot to check normality of CGPA for each Depression class
ggqqplot(train_data, x = "CGPA", facet.by = "Depression")
```

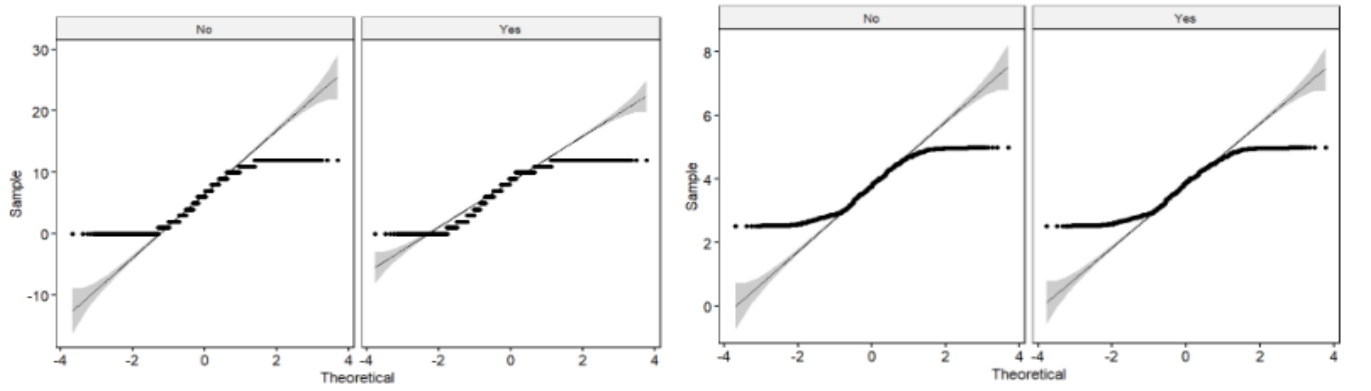

Figure 22 QQ Plot

4. Shapiro Test

The Shapiro-Wilk test is a statistical test that provides a formal assessment of normality. It calculates a test statistic (W) that measures how well the data conforms to a normal distribution. The test returns a p-value, which helps determine whether to reject the null hypothesis that the data is normally distributed.

- **W Statistic**: Values close to 1 suggest that the data is approximately normal.
- **P-Value** -
    - **High p-value (> 0.05)**: Fail to reject the null hypothesis (data is normally distributed).
    - **Low p-value (≤ 0.05)**: Reject the null hypothesis (data is not normally distributed).

Shapiro Test for Positive Values

```{r}

# Perform Shapiro-Wilk test
shapiro_results <- list()

for (i in numeric_vars) {
  shapiro_results[[i]] <- shapiro.test(stratified_sample_positive[[i]])
  cat("Shapiro-Wilk test for", i, ":\n")
  print(shapiro_results[[i]])
  cat("\n")
}
```

```
Shapiro-Wilk test for CGPA :

        Shapiro-Wilk normality test

data:  stratified_sample_positive[[i]]
W = 0.9478, p-value < 2.2e-16


Shapiro-Wilk test for Study_Hours :

        Shapiro-Wilk normality test

data:  stratified_sample_positive[[i]]
W = 0.90744, p-value < 2.2e-16
```

Shapiro Test for Negative Values

```{r}

# Perform Shapiro-Wilk test
shapiro_results <- list()

for (i in numeric_vars) {
  shapiro_results[[i]] <- shapiro.test(stratified_sample_negative[[i]])
  cat("Shapiro-Wilk test for", i, ":\n")
  print(shapiro_results[[i]])
  cat("\n")
}
```

```
Shapiro-Wilk test for CGPA :

        Shapiro-Wilk normality test

data:  stratified_sample_negative[[i]]
W = 0.9428, p-value < 2.2e-16


Shapiro-Wilk test for Study_Hours :

        Shapiro-Wilk normality test

data:  stratified_sample_negative[[i]]
W = 0.93109, p-value < 2.2e-16
```

Figure 23 Shapiro Test

From figure 23 both tests show very low p-values (less than 2.2e-16), meaning we reject the null hypothesis that the data is normally distributed. The W statistics also indicate some deviation from normality, with CGPA being slightly closer to normal than Study Hours. In summary, the data for both CGPA and Study Hours are not normally distributed.

```
**Fit LDA Model**
```{r}
lda_model <- lda(Depression ~., data = train_data)
print(lda_model)
```
```

```
Coefficients of linear discriminants:
                                        LD1
GenderMale                        0.005262606
Academic_Pressure2                0.521345486
Academic_Pressure3                1.044305659
Academic_Pressure4                1.480363663
Academic_Pressure5                1.750000475
CGPA                              0.038889221
Study_Satisfaction2              -0.213882398
Study_Satisfaction3              -0.207059475
Study_Satisfaction4              -0.380803628
Study_Satisfaction5              -0.531149232
Sleep_Duration7-8 hours           0.043590392
Sleep_DurationLess than 5 hours   0.180957958
Sleep_DurationMore than 8 hours  -0.156006945
Dietary_HabitsModerate            0.263832286
Dietary_HabitsUnhealthy           0.554353773
Suicidal_ThoughtsYes              1.530300654
Study_Hours                       0.058765293
Financial_stress2                 0.247284519
Financial_stress3                 0.628512738
Financial_stress4                 0.850114289
Financial_stress5                 1.151208619
Fam_Mental_HistYes                0.155195781
Age_Groupgroup_2                 -0.476742751
Age_Groupgroup_3                 -0.075482108
Age_Groupgroup_4                 -1.892746715
```

Figure 24 LDA Model

From the result of both the QQ plot and the Shapiro test we can see that the data is not normally distributed, Normality can be treated; it is possible to use log transformation. However, we are analyzing how the models behave first and in case the LDA model performs best then the data transformation will take place otherwise it will be ignored. Moving on to the model fitting, figure 24 shows the model fitting and the summary of the model. The model is fitted using the training data set. The results show the means and the coefficients of linear discriminants for all the features.

3.4.1 LDA Model Testing

To test and evaluate the model performance, the fitted model is then used to predict the classes in the test data and compared to the actual classes in the test data. Figure 25 below shows the code for testing and prediction and the confusion matrix.

```r
lda_predictions <- predict(lda_model, newdata = test_data)$class
```

```r
lda_conf_matrix <-table( Predicted = lda_predictions, Actual =test_data$Depression)
lda_conf_matrix
```

```
          Actual
 Predicted   No  Yes
       No  1123  231
       Yes  320 1807
```

Percentage of correct predictions

```r
lda_correct_predictions <- sum(diag(lda_conf_matrix)) / sum(lda_conf_matrix)
print(paste("Overall Correct Predictions Percentages: ", lda_correct_predictions))
```

```
 [1] "Overall Correct Predictions Percentages:  0.841712151680552"
```

Figure 25 LDA Test

From the results shown in figure 25, we conclude that:

- Out of the 3481 rows in the test dataset the model successfully predicted 2930 depression cases. Therefore 551 depression cases are false positive and false negative.
- The accuracy of this model is about 84.17%. In other words, 84.2% of the times, it predicts correctly.

3.5 Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis (QDA) is a classification technique that assumes different covariance matrices for each class, allowing for more flexibility compared to Linear Discriminant Analysis (LDA). QDA models the probability distribution of each class using a multivariate normal distribution and then classifies new observations based on the likelihood of belonging to each class. This method is particularly useful when the decision boundaries between classes are nonlinear. The primary advantage of QDA is its ability to model complex relationships between features, but it requires a larger dataset to estimate the covariance matrices accurately and may be less stable with small sample sizes or highly collinear features.

Same as LDA, QDA needs to meet some data assumption from the model to perform accurate. These assumptions were examined during the LDA analysis.

- Normality
- Linear independency
- Sufficient data

From the LDA analysis all assumptions were met. Moving on, for Quadratic Discriminant Analysis (QDA), the balance between variables is crucial. From the previous Exploratory Data Analysis (EDA), we gained insights into the data distribution and identified some important considerations regarding *Age Grouping*. Specifically, out of the four groups, two have minimal representation and hold almost no values compared to the other two groups. To maintain the integrity of the dataset, we will remove these two underrepresented groups. Additionally, as these groups are considered outliers, it is a good practice to eliminate them at this stage.

Figure 26 Age Group Removal

Figure 26 shows that new dataset dimensions being reduced after removing the students from the group ages 3 and 4. The picture also shows that there are only 2 groups in the age grouping column.

The next step for the QDA is to create a new dataset using stratification method this because QDA is sensible to the distribution, so it is important to keep in check the correct distribution of values for positive and negative cases.

```
**Get the stratification Sample**
```{r}
#Total size 27834, 75% = 20875 (10438, 10437) , 25% = 6958
set.seed(2025)
index_quadratic <- sampling:::strata(depression_dataset_copy, stratanames = c('Depression'), size =c(10438, 10437), method = 'srswor')
```

Separate and set the Stratification in TEsting and Training
```{r}
training_quadratic  <- depression_dataset_copy[index_quadratic$ID_unit,]
testing_quadratic <- depression_dataset_copy[- index_quadratic$ID_unit,]
```

New Datasets Dimensions
```{r}
# Training
dim(training_quadratic)
table(training_quadratic$Depression)


# Testing
#dim(testing_quadratic)
#table(testing_quadratic$Depression)
```
```

```
[1] 20875    12

    0     1
10437 10438
```

Figure 27 QDA Data Sample

Figure 27 shows the "new" datasets created for training and testing purposes for the QDA model. As the picture shows the distribution between the positive and negative cases are quite similar in numbers to have a good balance to meet the QDA requirements. The next step then is to load the QDA model with the training dataset (figure 28):

3.5.1 QDA Model Testing

```
**Fit the QDA Model**
```{r}
#quadratic_model<-
 qda(Depression ~. , data = training_quadratic)

```
```

```
Call:
qda(Depression ~ ., data = training_quadratic)

Prior probabilities of groups:
       0        1
0.499976 0.500024

Group means:
  GenderMale Academic_Pressure     CGPA Study_Satisfaction Sleep_Duration7-8 hours Sleep_DurationLess than 5 hours Sleep_DurationMore than 8 hours Dietary_HabitsModerate
0  0.5574399          2.364473 3.811924            3.218070               0.2587908                       0.2530421                       0.2576411              0.3776947
1  0.5573865          3.691224 3.845412            2.746982               0.2716996                       0.3247749                       0.1867216              0.3400077
  Dietary_HabitsUnhealthy Suicidal_ThoughtsYes Study_Hours Financial_stress Fam_Mental_HistYes Age_Groupgroup_2
0               0.2611862            0.3194405    6.231101         2.521510          0.4499377        0.4436141
1               0.4488408            0.8557195    7.765185         3.580284          0.5079517        0.2466948
```

Figure 28 QDA Model

The next step then is to load the QDA model with the training dataset to obtain the predicted values and compare with the real values, by doing this the prediction percentage is being calculated as previously done with the other models.

1. Fit the QDA Model

```{r}
quadratic_model<- qda(Depression ~. , data = training_quadratic)

```

After having the QDA model it is ready to test with test data:

2. Create a confusion matrix

```{r}
qda_predictions <- predict(quadratic_model, newdata = testing_quadratic)$class
qda_conf_matrix <-table( Predicted = qda_predictions, Actual =testing_quadratic$Depression)
qda_conf_matrix
```

```
          Actual
 Predicted    0    1
         0  916  974
         1  178 4891
```

3. Percentage of correct predictions

```{r}
qda_correct_predictions <- sum(diag(qda_conf_matrix)) / sum(qda_conf_matrix)
print(paste("Overall Correct Predictions Percentages: ", qda_correct_predictions))
```

```
[1] "Overall Correct Predictions Percentages:  0.834458973990516"
```

Figure 29 QLD Test

From figure 29 the results, we conclude that:

- Out of the 6959 rows in the test dataset the model successfully predicted 5807 depression cases. Therefore 1152 depression cases are false positive and false negative.
- The accuracy of this model is about 83.44%. In other words, 83.4% of the times, it predicts correctly.

3.6 Decision Tree Model

A tree model, also known as a decision tree, is a machine learning algorithm used for classification and regression tasks. It works by recursively splitting the data into subsets based on the values of input features, creating a tree-like structure of decisions. Each internal node represents a decision based on a feature, each branch represents the outcome of the decision, and each leaf node represents a predicted outcome or class. The main advantages of tree models are their interpretability and simplicity, as they mimic human decision-making processes. They are also non-parametric, meaning they do not assume any underlying distribution of the data, making them versatile and effective for various types of data.

Another relevant factor about decision trees unlike some other models, decision trees are non-parametric, meaning they do not assume any specific distribution for the data. They can handle both numerical and categorical features without needing to standardize or normalize them. Therefore, are a great tool when the distribution or data cleaning process is not the best. In this case the cleaning process has been done but it does not represent relevance to the model. Same as the previous models we have fit the model using the training data to verify its accuracy percentage.

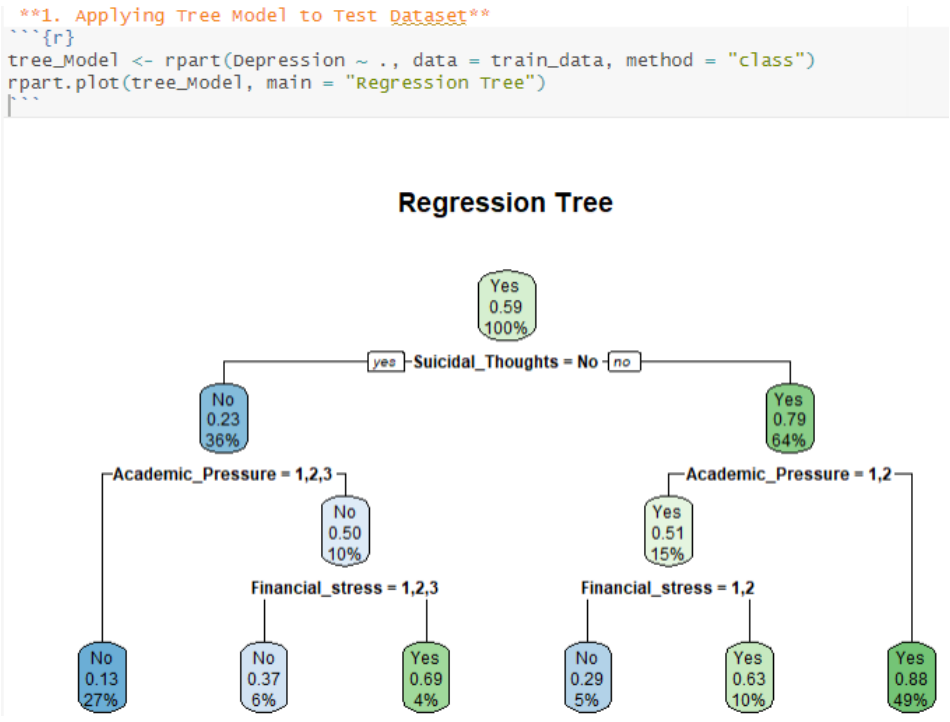

Figure 30 Tree Model

From the model representation [Figure 30] it is possible to notice that there are 3 main categories that the model uses to generate the decisions. This aligns to some of our assumption in the EDA Summary section.

- Suicidal Thoughts
- Academic Pressure
- Financial Stress

3.6.1 Decision Tree Model Testing

The next step then is to load the Tree Model with the training dataset to obtain the predicted values and compare with the real values, by doing this the prediction percentage is being calculated as previously done with the other models.

```
**2. Create a confusion matrix**
```{r}
tree_predictions<-predict(tree_Model,test_data,type = "class")# class menas classification
tree_conf_matrix <-table(tree_predictions,test_data$Depression)
tree_conf_matrix

 # Extract values from Confusion Matrix
TN <- tree_conf_matrix[1, 1]  # True Negatives
FP <- tree_conf_matrix[1, 2]  # False Positives
FN <- tree_conf_matrix[2, 1]  # False Negatives
TP <- tree_conf_matrix[2, 2]  # True Positives

# Compute Performance Metrics
accuracy <- (TP + TN) / (TP + TN + FP + FN)
precision <- TP / (TP + FP)
recall <- TP / (TP + FN)
f1_score <- 2 * ((precision * recall) / (precision + recall))

# Print Metrics
cat("Decision Tree Model Performance Metrics:\n")
cat("Model Accuracy:", round(accuracy, 4), "\n")
cat("Precision:", round(precision, 4), "\n")
cat("Recall:", round(recall, 4), "\n")
cat("F1 Score:", round(f1_score, 4), "\n")

```
```

```
tree_predictions   No  Yes
              No  1072  264
              Yes  371 1774
Decision Tree Model Performance Metrics:
Model Accuracy: 0.8176
Precision: 0.8705
Recall: 0.827
F1 Score: 0.8482
```

Figure 31 Tree Model Test

From the results in figure 31, we conclude that:

- Out of the 3481 rows in the test dataset the model successfully predicted 2846 depression cases. Therefore 635 depression cases are false positive and false negative.
- The accuracy of this model is about 81.75%. In other words, 81.7% of the times, it predicts correctly.

3.7 Random Forest Algorithm

A Random Forest is an ensemble learning method used for classification and regression tasks. It operates by constructing multiple decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. This approach enhances the accuracy and robustness of the model by reducing overfitting.

The output of a Random Forest model includes the predicted values and an assessment of variable importance. The variable importance plot shows which features contribute the most to the prediction, providing insights into the underlying patterns in the data. In addition to that, same as Tree model, the do not require stringent assumptions about the data, such as normality or linear relationships. It can handle both numerical and categorical features, manage missing values, and is resistant to overfitting due to its ensemble nature.

In order to have a correct distribution of the data, it is necessary to split the data in test and train sets before fitting/ training the model (figure 32):

```
**Splitting the data**
```{r}
# Ensure 'Depression' is a factor
stratified_sample$Depression <- as.factor(stratified_sample$Depression)

# Split data using stratified sampling (75% training, 25% testing)
set.seed(2025)
trainIndex <- createDataPartition(stratified_sample$Depression, p = 0.75, list = FALSE)
train_data <- stratified_sample[trainIndex, ]
test_data <- stratified_sample[-trainIndex, ]
```
```

Figure 32 Random Forest Splitting Dataset

3.7.1 Random Forest Model Testing

Once the sampling is done the next step is fit/ train the model (figure33).

```
**Fit model**
```{r}
# Train Random Forest Model
set.seed(2025)
rf_model <- randomForest(Depression ~ ., data = train_data, ntree = 500, mtry = 3, importance = TRUE)

# Print Model Summary
print(rf_model)

# Make Predictions
rf_predictions <- predict(rf_model, newdata = test_data)

# Ensure factor levels match
rf_predictions <- factor(rf_predictions, levels = levels(test_data$Depression))

# Compute Confusion Matrix
conf_matrix <- confusionMatrix(rf_predictions, test_data$Depression)
print(conf_matrix)

```
```

Figure 33 Random Forest Model

The result of the random forest model is shown in the figure 34:

```
```{r}
# Extract values from Confusion Matrix
TN <- conf_matrix$table[1, 1]  # True Negatives
FP <- conf_matrix$table[1, 2]  # False Positives
FN <- conf_matrix$table[2, 1]  # False Negatives
TP <- conf_matrix$table[2, 2]  # True Positives

# Compute Performance Metrics
accuracy <- (TP + TN) / (TP + TN + FP + FN)
precision <- TP / (TP + FP)
recall <- TP / (TP + FN)
f1_score <- 2 * ((precision * recall) / (precision + recall))

# Print Metrics
cat("Model Accuracy:", round(accuracy, 4), "\n")
cat("Precision:", round(precision, 4), "\n")
cat("Recall:", round(recall, 4), "\n")
cat("F1 Score:", round(f1_score, 4), "\n")
```
```

```
Model Accuracy: 0.8388
Precision: 0.8665
Recall: 0.8594
F1 Score: 0.8629
```

Figure 34 Random Forest Model Test

From the results in figure 34, we conclude that:

- Out of the 3481 rows in the test dataset the model successfully predicted 2920 depression cases. Therefore 561 depression cases are false positive and false negative.
- The accuracy of this model is about 83.88%. In other words, 83.9% of the times, it predicts correctly.

4. Model Performance

In the field of mental health analytics, **predicting depression** based on various psychological, social, and lifestyle factors is a crucial challenge. Machine learning models provide a powerful approach to analyzing complex patterns in depression-related data. However, selecting the **best model** requires evaluating multiple algorithms based on key performance metrics such as **accuracy, precision, recall, and F1-score**.

This study compares five machine learning models—**Logistic Regression, Random Forest, Decision Tree, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA)**—to determine which model is best suited for depression prediction. Each model is trained and tested using a structured dataset containing features such as **academic pressure, sleep duration, dietary habits, suicidal thoughts, study hours, financial stress, and family mental health history.**

4.1 Accuracy, Precision, Recall and F1 Score:

The performance of each model is evaluated using four critical metrics:

- **Accuracy**: Measures the overall correctness of the model.
- **Precision**: Evaluates how many predicted positive cases are actually positive.
- **Recall (Sensitivity)**: Measures the model's ability to identify true positive cases.
- **F1-Score**: Balances precision and recall to assess the model's reliability.

By comparing these metrics, we aim to determine which model provides the **most reliable predictions** for identifying individuals at risk of depression. The findings from this analysis will help in **selecting an optimal model** for real-world applications in mental health screening and early intervention. The table below presents the **accuracy, precision, recall, and F1-score** of each model used in the study.

Table 4 Model's Performance

Model	Accuracy	Precision	Recall	F1_Score
Logistic Model	0.8440	0.8808	0.8568	0.8686
Reduce Logistic Model	0.8429	0.8783	0.8569	0.8675

LDA Model	0.8417	0.8867	0.8496	0.8678
QDA Model	0.8345	0.8339	0.9649	0.8946
Random Forest Model	0.8388	0.8665	0.8594	0.8629
Decision Tree Model	0.8176	0.8705	0.8270	0.8482

Based on the results in table 4:

- The **Logistic Model** stands out with the highest accuracy (0.8440) and a solid F1-Score (0.8686). It demonstrates a good balance between precision (0.8808) and recall (0.8568), indicating its overall reliability in predicting depression in this dataset.
- The **Reduce Logistic Model** and **LDA** also perform well, with comparable accuracy, precision, recall, and F1-scores.
- The **QDA** model has the highest recall (0.9649) and F1-score (0.8946), which suggests it is particularly effective in identifying true positives but may suffer from lower precision (0.8339).
- **Random Forest** and **Decision Tree** models show slightly lower accuracy, and F1-scores compared to the logistic models and LDA but are still viable options for depression prediction with balanced performance metrics.

Overall, the **Logistic Model** is the most balanced and accurate, but the **Reduce Logistic Model** and **LDA** are strong alternatives. The **QDA** model excels in recall, making it a good choice if the priority is to minimize false negatives. **Quadratic Discriminant Analysis (QDA)** shows good recall, it struggles with precision, making it less optimal for ensuring accurate positive classifications. **Decision Tree**, on the other hand, performs the worst due to its low recall and F1-score, leading to a higher rate of misclassification.

These findings suggest that **LDA is the most suitable model for practical applications in mental health prediction**, providing a strong foundation for decision-making in early detection and intervention strategies. However, further improvements such as **feature selection, hyperparameter tuning, or ensemble methods** could enhance model performance, leading to more refined and accurate depression predictions.

4.2 Confusion Matrices for Each Model

In terms of individual matrix elements, **true negatives (TN)** represent correctly predicted non-depressed individuals, while **false positives (FP)** indicate those incorrectly classified as depressed. **False negatives (FN)** are actual cases of depression that the model failed to detect, and **true positives (TP)** are those correctly identified as depressed individuals.

Table 5 Model's Confusion Matrix

Model	True Negative	False Positive	False Negative	True Positive
Logistic	1143	243	300	1795
Reduce Logistic	1144	248	299	1790
LDA	1123	231	320	1807
QDA	916	974	178	4891
Random Forest	1154	272	289	1766
Decision Tree	1072	264	371	1774

Based on the results from table 5,

- The QDA model stands out with a very high true positive count (4891) and the lowest false negative count (178). This makes it particularly effective in identifying true cases of depression, despite having a high false positive count (974).
- The LDA model also shows a good balance with a strong true positive count (1807) and a relatively low false positive count (231), making it another suitable choice for depression prediction.
- Logistic Regression and Reduce Logistic models provide a solid performance with high true negative counts (1143 and 1144 respectively) and moderate false negative counts.
- The Random Forest model demonstrates decent performance with a good true negative count (1154) and a moderate false negative count (289).
- Decision Tree shows the lowest performance among the models, with a higher false negative count (371) and the lowest true positive count (1774), indicating that it struggles to capture many true cases of depression.

Overall, while QDA excels in identifying true positives, LDA maintains a good balance between minimizing false negatives and maintaining a strong true positive count. Logistic Regression and Reduce Logistic models are reliable alternatives, whereas the Decision Tree model performs the worst in this context.

Conclusion:

In conclusion the dataset yields several key insights into the factors influencing student mental health. Among the predictive models tested, the Logistic Model, Reduced Logistic Model, and Linear Discriminant Analysis (LDA) demonstrated strong predictive performance, with the Logistic Model achieving the highest accuracy. The Quadratic Discriminant Analysis (QDA) model was particularly effective in identifying true cases of depression due to its high recall rate. Our analysis highlighted three primary factors that significantly contributed to depression predictions: suicidal thoughts, academic pressure, and financial stress. Students exhibiting suicidal ideation, experiencing high academic pressure, or facing financial instability were at the highest risk of expressing depression. Additionally, other variables, including unhealthy dietary habits, insufficient sleep, age, gender, study hours, and study satisfaction, also played a role in predicting depression. The data indicated that younger students, male students, and those experiencing high academic or financial stress were more vulnerable to depression.

However, a key limitation of this study is that the linearity assumption was not fully met, which may affect the interpretation of some model results. Future research should explore non-linear modeling approaches to enhance predictive accuracy. Despite this limitation, our findings provide valuable insights into the multifaceted nature of student mental health and underscore the importance of targeted interventions to mitigate depression risks. While further research is necessary for a more comprehensive understanding, these findings suggest that mental health interventions tailored to high-risk groups particularly those struggling with financial stress, academic pressure, or suicidal ideation could be instrumental in effectively managing and reducing depression among students.

References:

1. Cleveland Clinic. (n.d.). *Depression: Symptoms, causes, treatment*. Retrieved from https://my.clevelandclinic.org/health/diseases/9290-depression
2. NIH National Library of Medicine. (n.d.). *The epidemiology of depression*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC3330161/
3. Kaggle. (n.d.). *Depression dataset*. Retrieved from Student Depression Dataset.
4. Kaggle. (n.d.). hopesb. Retrieve from Shodolamu Opeyemi | Contributor | Kaggle
5. Graves, B. S., Hall, M. E., Dias-Karch, C., Haischer, M. H., & Apter, C. (2021). Gender differences in perceived stress and coping among college students. *PLOS ONE, 16*(8), e0255634. https://doi.org/10.1371/journal.pone.0255634
6. Liu, Y., & Wang, Z. (2024). Depression, anxiety, and student satisfaction with university life: A longitudinal study. *Humanities and Social Sciences Communications, 11*, Article 286. https://doi.org/10.1038/s41599-024-03686-y
7. McDaid D, Park AL. The economic case for investing in the prevention of mental health conditions in the UK. Care Policy and Evaluation Centre, Department of Health Policy, London School of Economics and Political Science, London; 2022.
8. J.R. Williamson, *et al.*Detecting depression using vocal, facial and semantic communication cues AVEC 2016 - Proc. 6th Int. Work. Audio/Visual Emot. Challenge, Co-located with ACM Multimed (2016), pp. 11-18, 10.1145/2988257.2988263 2016.