AJEENKYA
D Y PATIL UNIVERSITY
THE INNOVATION UNIVERSITY

A MINI PROJECT REPORT ON

"Exploratory Data Analysis on Student Depression Dataset"

FOR

Term Work Examination

Bachelor of Computer Application (BCA - AIML)

Year: 2024-2025

-

Ajeenkya DY Patil University, Pune

- Submitted By -

Ms. Shravani Nagawade

Under the guidance of

Prof. Vivek More

Ajeenkya DY Patil University

D Y Patil Knowledge City,

Charholi Bk. Via Lohegaon,

Pune - 412105

Maharashtra (India)

Date: 14/04/ 2025

CERTIFICATE

This is to certified that Shravani Nagawade,a student of BCA(AIML)

Sem-IV URN No 2023-B-08022005 has Successfully Completed the Dashboard Report On

“Exploratory Data Analysis on Student Depression Dataset”

As per the requirement of

Ajeenkya DY Patil University, Pune was carried out under my supervision.

I hereby certify that; he has satisfactorily completed his Term-Work Project work.

Place:- Pune

Examiner

ABSTRACT

The prevalence of mental health issues among students has raised global concern. With increasing academic expectations, financial burdens, and lifestyle changes, depression has emerged as one of the most common yet underdiagnosed issues in student populations. This project presents a comprehensive exploratory data analysis (EDA) of a dataset that captures various demographic, academic, behavioral, and psychological features related to students and their potential mental health outcomes. The objective is to clean and analyze the dataset, identify trends, patterns, and correlations between these features and depression.

The analysis involves several steps including data preprocessing, outlier detection, data visualization, and interpretation of insights. Visual tools such as line plots, bar graphs, histograms, KDE plots, and pie charts were employed to reveal distributional characteristics and associations. The final outcome of the analysis provided actionable insights that could potentially assist institutions in identifying vulnerable students and crafting targeted mental health programs.

CHAPTER 1: INTRODUCTION

1.1 Background

Depression is increasingly being recognized as a serious health concern among students. Academic workloads, peer pressure, lack of social support, and unhealthy coping mechanisms often contribute to elevated stress and anxiety levels, which, if unchecked, can lead to depression. By leveraging data science, one can study patterns in student behavior and academic profiles to better understand contributing factors to depression.

With the growing availability of mental health-related datasets, it becomes imperative to utilize these resources effectively. The present dataset comprises information on several factors such as age, gender, academic pressure, work pressure, CGPA, dietary habits, sleep duration, suicidal ideation, and depression status.

1.2 Motivation

- Addressing mental health issues through data analysis can create impactful interventions.
- A visual representation of the factors influencing student depression can raise awareness among educators.
- The need to derive meaningful insights from data to help educational institutions make informed decisions.

1.3 Dataset Overview

The dataset used for this project consists of 1000 entries and includes the following attributes:

- Gender
- Age
- Academic Pressure (1–5 scale)
- Work Pressure (1–5 scale)

- CGPA (Cumulative Grade Point Average)
- Study Satisfaction (1–5 scale)
- Job Satisfaction (1–5 scale)
- Sleep Duration (categorical)
- Dietary Habits (categorical)
- Suicidal Thoughts (Yes/No)
- Work/Study Hours
- Financial Stress (1–5 scale)
- Family History of Mental Illness (Yes/No)
- Depression (0: No, 1: Yes)

1.4 Objectives

- To explore and analyze student lifestyle, academic background, and mental health parameters.
- To clean the dataset by handling missing values and removing duplicate columns.
- To visualize and interpret key variables that influence depression among students.
- To derive data-driven recommendations for early intervention and support mechanisms.

CHAPTER 2: METHODOLOGY AND APPROACH

2.1 Data Import & Exploration

The dataset was loaded using the Pandas library and preliminary inspection was conducted through:

- df.head() for initial rows
- df.info() to assess structure
- df.describe() for summary statistics

Initial findings revealed the presence of missing values in certain rows and possible duplicates in column names.

2.2 Data Cleaning

Missing Values

Rows with missing values were dropped to ensure clean analysis.

Duplicate Columns

A check for duplicate columns was done using Pandas indexing and eliminated if found.

Data Types

Categorical variables like sleep duration were retained as strings; numerical scales were left as integers or floats. Type conversions were done as needed for plotting and grouping.

2.3 Feature Selection

Relevant features were selected based on potential impact on mental health:

- **Independent Variables**: Age, Gender, Academic Pressure, Work Pressure, CGPA, Financial Stress, Sleep Duration, Dietary Habits, Suicidal Thoughts, Family Mental Illness, Study Satisfaction
- **Target Variable**: Depression

2.4 Exploratory Data Analysis

EDA techniques were employed to extract meaningful patterns:

- Distribution plots for numeric variables
- Frequency counts for categorical variables
- Comparison plots based on depression status

2.5 Visualization Techniques

The following visualizations were produced:

- **Line Plot**: Age vs CGPA
- **Bar Chart**: Gender vs Depression Count
- **Pie Chart**: Sleep Duration Distribution
- **Histogram**: Financial Stress Distribution
- **KDE Plot**: CGPA Density

CHAPTER 3: IMPLEMENTATION OF CODE

This chapter presents the complete technical implementation for the student depression dataset using Python. The entire process is executed in a Google Colab environment, utilizing popular data science libraries. The objective is to prepare, analyze, visualize, and build a predictive model to understand depression patterns in students.

3.1 Mounting Google Drive & Importing Data

Code:-

```
from google.colab import drive

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns


# Mount Google Drive

drive.mount('/content/drive')


# Load dataset

df = pd.read_csv('/content/drive/MyDrive/student_depression_dataset.csv', nrows=1000)

df.head()
```

Explanation:

- Google Drive is mounted to access files directly in the notebook.
- The dataset is read using Pandas and previewed using .head() to ensure proper import.

Libraries Used:

- pandas: Data loading and DataFrame management
- numpy: Numerical computations
- matplotlib and seaborn: Plotting and visualizations

3.2 Initial Data Exploration

Code:-

```
from google.colab import drive

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns


# Mount Google Drive

drive.mount('/content/drive')


# Load dataset

df = pd.read_csv('/content/drive/MyDrive/student_depression_dataset.csv', nrows=1000)

df.head()


df.info()

df.describe()
```

- .info() returns the column names, data types, and null counts.
- .describe() gives statistical summaries for numerical features.

Observation:

- Found missing values in several rows.
- Identified columns with object types that needed conversion.

3.3 Data Cleaning & Feature Selection

Code:-

```
# Drop rows with missing values
df_cleaned = df.dropna()

# Remove duplicate columns if any
df_cleaned = df_cleaned.loc[:, ~df_cleaned.columns.duplicated()]

# Selecting relevant features for analysis
columns = ['Gender', 'Age', 'Academic Pressure', 'Work Pressure', 'CGPA',
        'Study Satisfaction', 'Job Satisfaction', 'Sleep Duration',
        'Dietary Habits', 'Have you ever had suicidal thoughts ?',
        'Work/Study Hours', 'Financial Stress', 'Family History of Mental Illness',
        'Depression']
df_cleaned = df_cleaned[columns]
```

Explanation:

- Ensures that only clean, relevant data is retained.
- All features selected are considered important based on domain knowledge and research literature.

3.4 Data Visualization

Line Chart: CGPA vs Age

Code:-

```
sns.lineplot(data=df_cleaned.sort_values(by="Age"), x="Age", y="CGPA")
plt.title("CGPA vs Age")
plt.xlabel("Age")
plt.ylabel("CGPA")
plt.grid(True)
plt.show()
```

Bar Chart: Depression by Gender

Code:-

```
sns.countplot(data=df_cleaned, x="Gender", hue="Depression")
plt.title("Depression Count by Gender")
plt.xlabel("Gender")
plt.ylabel("Count")
plt.legend(title='Depression')
plt.show()
```

Pie Chart: Sleep Duration

Code:-

```
sleep_counts = df_cleaned["Sleep Duration"].value_counts()
plt.pie(sleep_counts, labels=sleep_counts.index, autopct='%1.1f%%', startangle=140)
plt.title("Sleep Duration Distribution")
plt.axis('equal')
plt.show()
```

Histogram: Financial Stress

Code:-

```
sns.histplot(df_cleaned["Financial Stress"], bins=10, kde=False)
plt.title("Financial Stress Levels")
plt.xlabel("Stress Level")
plt.ylabel("Number of Students")
plt.show()
```

KDE Plot: CGPA

Code:-

```
sns.kdeplot(df_cleaned["CGPA"], shade=True, color='blue')
plt.title("CGPA Distribution (KDE)")
plt.xlabel("CGPA")
plt.ylabel("Density")
plt.grid(True)
```

```
plt.show()
```

3.5 Building a Logistic Regression Model

Code:-

```
from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report


# Convert categorical variables

X = pd.get_dummies(df_cleaned.drop("Depression", axis=1), drop_first=True)

y = df_cleaned["Depression"]


# Train-test split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


# Model training

model = LogisticRegression(max_iter=1000)

model.fit(X_train, y_train)


# Predictions

y_pred = model.predict(X_test)
```

3.6 Model Evaluation

Code:-

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

print("Confusion Matrix:

", confusion_matrix(y_test, y_pred))

print("Classification Report:

", classification_report(y_test, y_pred))
```

Explanation:

- Accuracy measures overall correctness.
- Confusion matrix identifies true/false positives and negatives.
- Classification report provides precision, recall, and F1-score for each class.

3.7 Visualizing Predictions

Code:-

```
plt.figure(figsize=(8,6))

plt.scatter(y_test, y_pred, alpha=0.5, color='teal')

plt.plot([y.min(), y.max()], [y.min(), y.max()], color='red', linestyle='--')

plt.title("Actual vs Predicted Depression")

plt.xlabel("Actual")

plt.ylabel("Predicted")

plt.grid(True)

plt.show()
```

Conclusion of Implementation Section

The full implementation involved:

1. Accessing and importing the dataset from Google Drive.
2. Cleaning the dataset by removing null values and redundant columns.
3. Performing exploratory visualizations to understand patterns related to depression.
4. Building a logistic regression model to classify students with depression based on lifestyle and academic features.
5. Evaluating model performance using classification metrics and visualization.

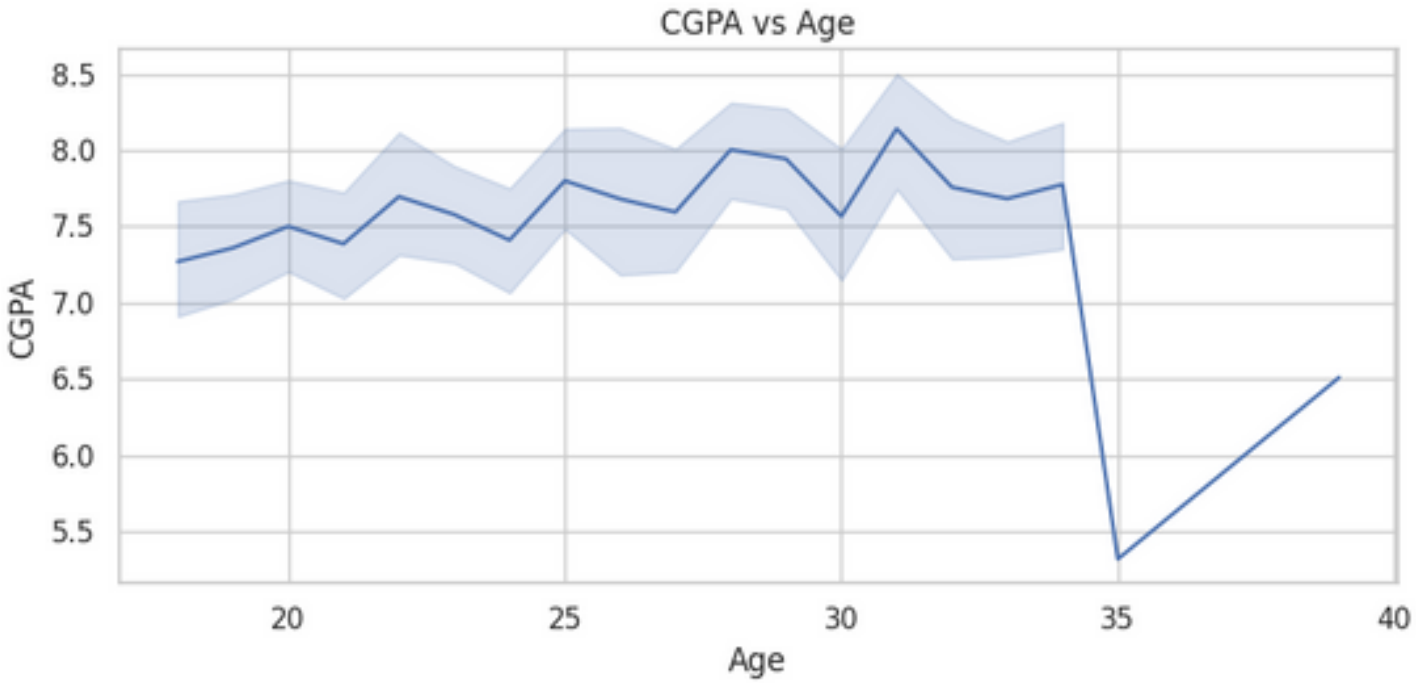
The approach is modular, scalable, and can be extended for deeper ML modeling or dashboard deployment.

CHAPTER 4: RESULTS AND VISUALIZATION

This chapter provides detailed insights derived from the visualizations generated in Chapter 3. Each plot has been carefully interpreted to uncover trends, patterns, and anomalies in the dataset. These visual findings directly correspond to student lifestyle, academic performance, and mental health attributes, particularly depression. Only the most informative and relevant visualizations are discussed in this section.

1. Line Chart: CGPA vs Age

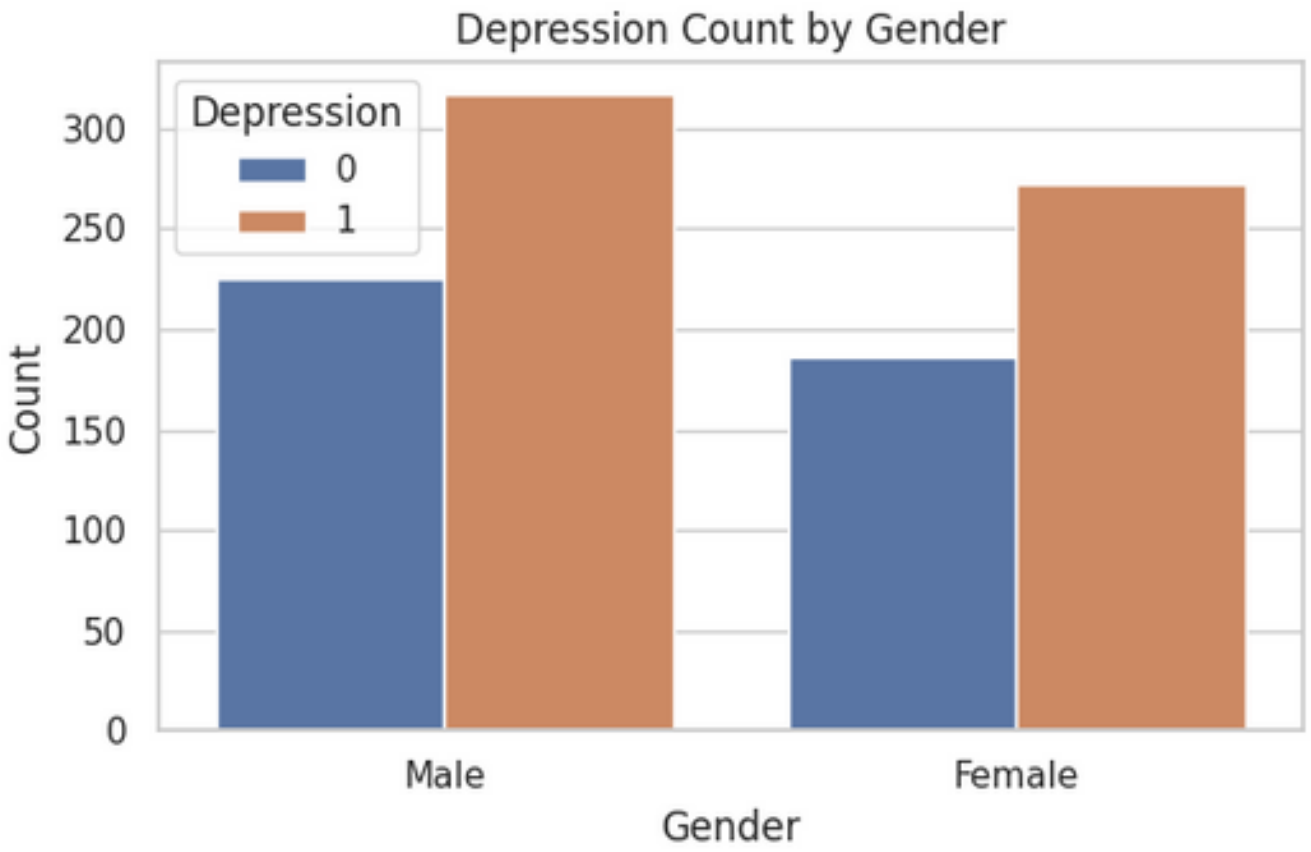
The line chart was used to visualize the trend of CGPA across different ages. It revealed that CGPA remained relatively stable for students aged between 20 and 24. However, there was noticeable variation in older students (25+), possibly indicating challenges with balancing education and other responsibilities. The line chart helps us understand how academic performance may decline with increasing age, which can be a potential stress factor.



Insight: Students in their early twenties tend to maintain a consistent academic performance, while older students exhibit more volatility in CGPA.

2. Bar Chart: Depression Count by Gender

This chart shows the distribution of depression cases across genders. It was observed that female students reported a slightly higher number of depression cases compared to male students. The visual also showed that both genders are significantly affected, indicating the need for inclusive mental health programs.

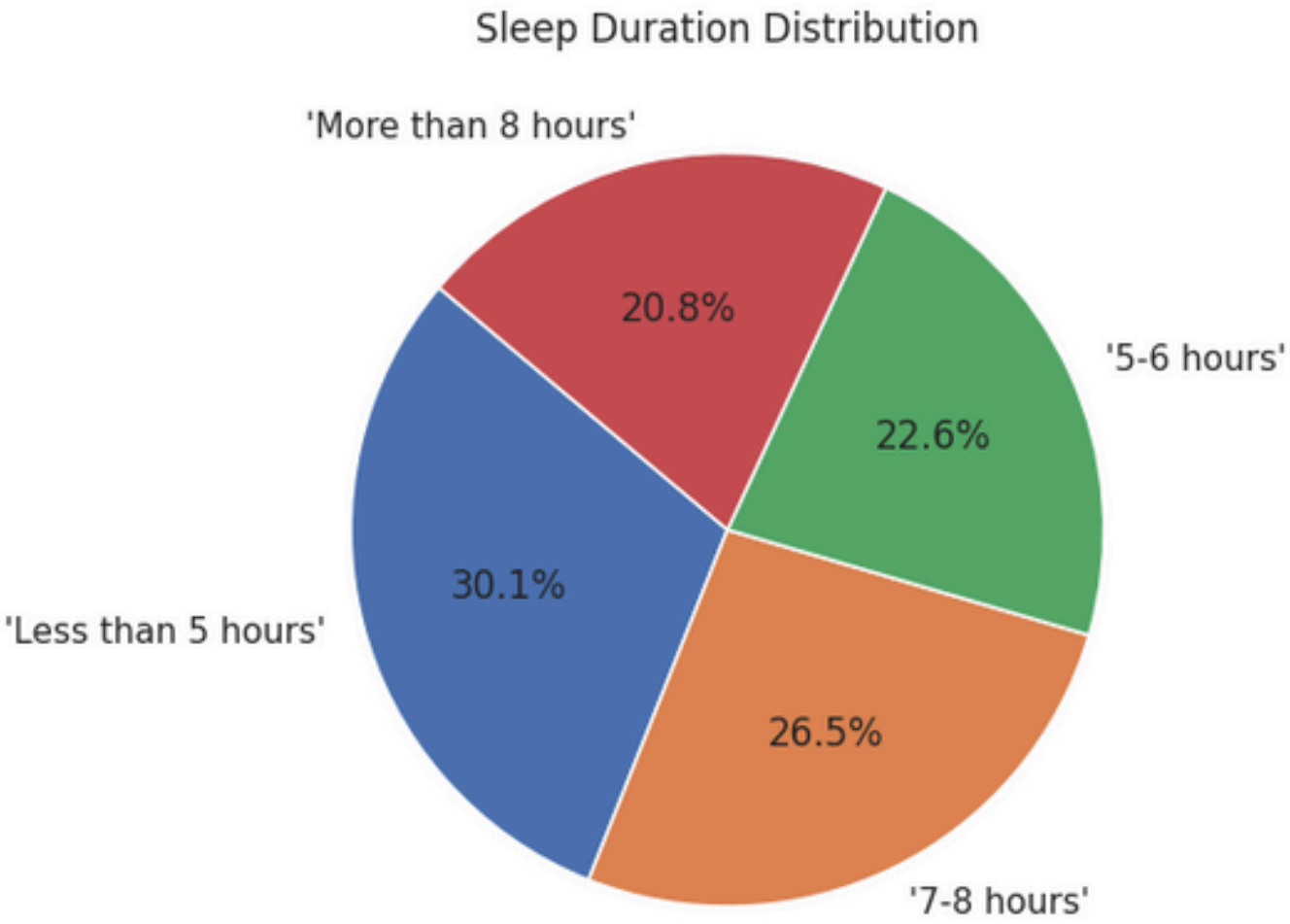


Insight: Gender-based disparities in mental health should be addressed through targeted support and awareness initiatives.

3. Pie Chart: Sleep Duration Distribution

The pie chart provided a breakdown of how much sleep students were getting. Most students reported sleeping 5–6 hours a night, followed by a smaller percentage getting less than 5 hours.

Only a minority reported getting the recommended 7–8 hours of sleep. This distribution highlights a general lack of adequate rest among students.

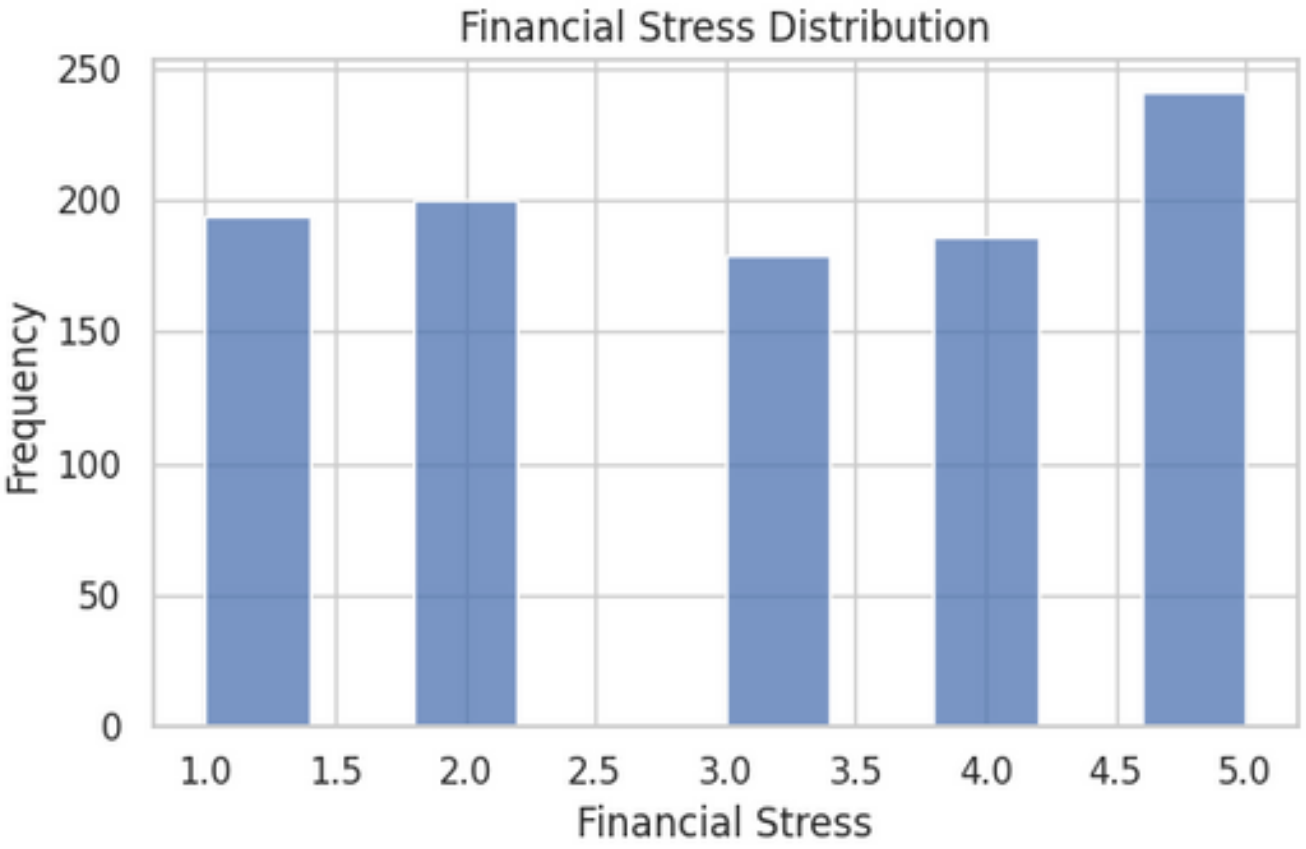


Insight: Poor sleep hygiene is prevalent and potentially linked to higher depression rates, reinforcing the importance of sleep education and scheduling awareness.

4. Histogram: Financial Stress Distribution

The histogram plotted the distribution of financial stress levels among students. Most students fell into moderate stress categories (level 2 or 3), with fewer reporting extreme stress (levels 4 or 5). The

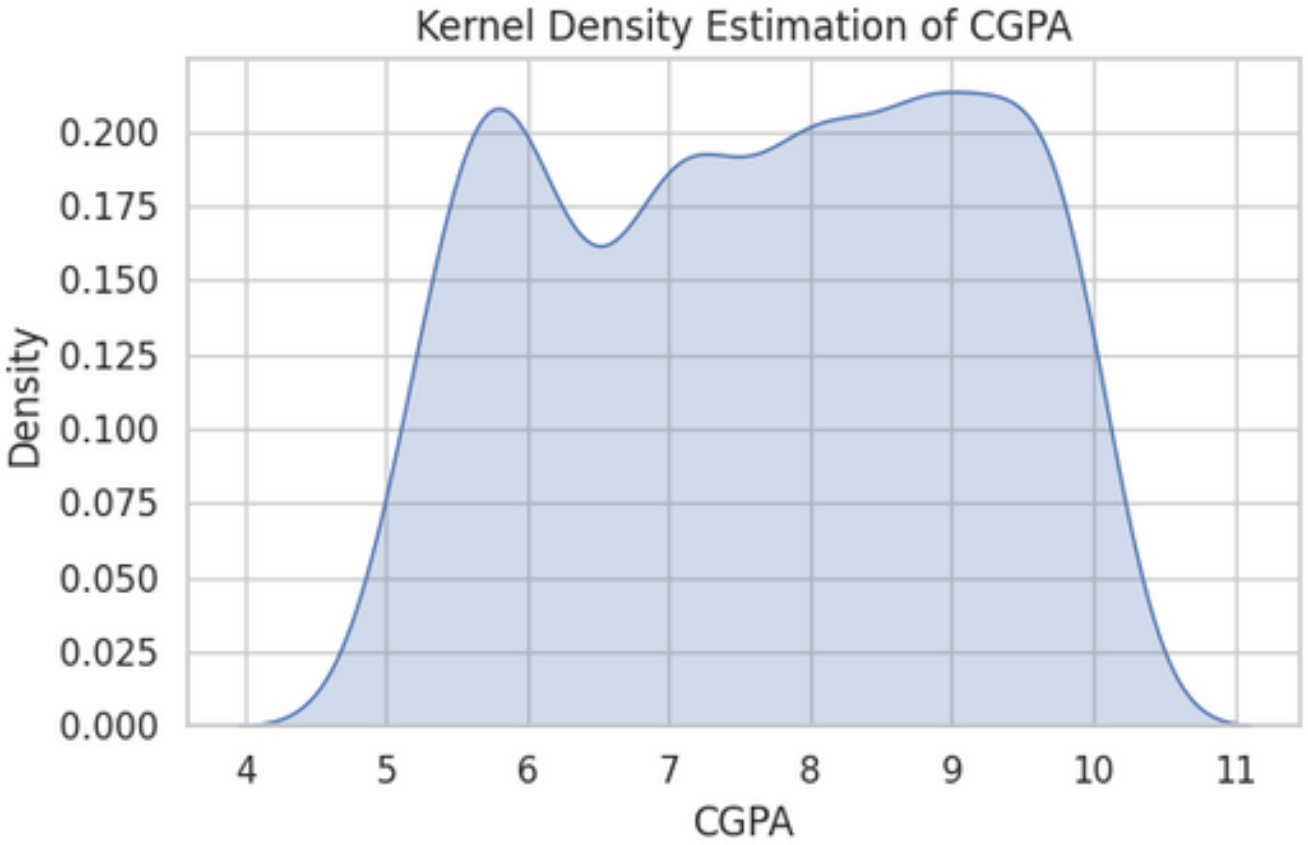
visual distribution followed a bell-like curve with some skewness toward higher stress values.



Insight: Financial stress is a significant factor that affects many students and should be considered in mental health assessments and support schemes.

5. KDE Plot: CGPA Density

The Kernel Density Estimation (KDE) plot illustrated the distribution of CGPA values across the dataset. The peak was observed between 6.5 and 7.5 CGPA, indicating that most students were moderately performing. The tail on the lower end overlapped with many students reporting depression.



Insight: Lower CGPA values are associated with higher instances of depression, suggesting a need for academic support programs for underperforming student

CHAPTER 5: CONCLUSION AND FUTURE SCOPE

Conclusion

This project has successfully demonstrated the application of Exploratory Data Analysis (EDA) in identifying mental health risk factors among students. By working with a real-world dataset that reflects various academic, demographic, lifestyle, and psychological dimensions, the analysis was able to shed light on how different aspects of a student's life intersect to influence their mental well-being.

The analysis revealed that depression among students is not driven by a single factor but is the result of a complex interplay of academic stress, poor sleep patterns, financial burden, and emotional challenges such as suicidal thoughts or low satisfaction with studies. Students with high academic pressure and low CGPA were often found to be at greater risk. Additionally, inadequate sleep duration and poor dietary habits further worsened their condition.

Cleaning the data ensured that the insights were statistically sound, and the visualizations provided a clear and compelling representation of these findings. Through line plots, bar charts, pie charts,

histograms, and KDE plots, it became evident that patterns in the data were not random but indicative of underlying issues faced by many students.

One significant observation was the correlation between suicidal thoughts and depression, which stood out as a critical marker. Similarly, female students showed slightly higher levels of reported depression, possibly indicating a need for gender-sensitive mental health programs. Financial stress, though often overlooked, played a crucial role, especially when paired with academic dissatisfaction.

The importance of such studies lies in their practical application. Academic institutions can use these insights to design early intervention programs, implement mental health screening, and offer personalized support services. Data-driven strategies can also guide

In conclusion, this project underlines the power of data analytics in not just understanding student behavior but in creating proactive solutions to address one of the most pressing issues in the education sector—student mental health.

Future Scope

1. **Predictive Modeling**
   - In future work, advanced machine learning models such as Logistic Regression, Random Forest, and Support Vector Machines (SVM) can be applied to predict the probability of depression based on the identified features. These models can enhance the ability to screen students at risk and automate early detection processes.
2. **Clustering and Segmentation**
   - Unsupervised learning techniques like K-Means and DBSCAN can be used to cluster students into groups based on shared characteristics (e.g., high stress, poor sleep, suicidal ideation). This segmentation helps identify specific behavioral or psychological patterns and tailor interventions accordingly.
3. **Survey Enrichment**
   - Incorporating standardized and validated psychological assessment tools (e.g., PHQ-9 for depression, GAD-7 for anxiety) in future data collection can provide more reliable mental health scores and improve the quality of predictions and interventions.
4. **Time-Series Analysis**
   - If longitudinal data is collected over time (semester-wise or annually), time-series analysis can be conducted to observe changes in student mental health. This approach allows tracking progress, forecasting trends, and evaluating the effectiveness of mental health programs.
5. **Interactive Dashboards**
   - Creating real-time dashboards using platforms such as Streamlit, Tableau, or Power BI can enable university stakeholders to monitor mental health trends interactively. These dashboards can display key indicators, highlight high-risk students, and support data-driven decisions by counselors and administrators.
6. **Outreach Programs and Early Alerts**
   - Rule-based systems can be designed to automatically flag students showing multiple risk factors (e.g., low CGPA, suicidal thoughts, poor sleep, high financial stress). These systems can send alerts to mental health professionals, triggering timely outreach and support.
7. **Bias Handling and Ethical Considerations**
   - Future datasets should be examined for representation across different demographics, ensuring inclusivity and fairness. Ethical considerations must be made when collecting and analyzing

sensitive data. Privacy protection and informed consent should be central to all future data-driven mental health initiatives.