DATA ANALYSIS

**TOPIC:**

***Student Depression***

**SECTION:**

**BS INFOTECH – 3B**

**MEMBERS:**

Elizabeth Tarrayo

Angeline Doinog

Jancee Gylle Lim

Juan Floren Boncayao

Mat koski Monares

Christian Arman Opanes

Dave Galicia

**SUBMITTED TO:**

**Ms. Myka Pilpa**

**Instructor**

**INTRODUCTION**

This project thoroughly investigates the various factors that influence student depression, using a dataset titled "Student Depression Dataset." The dataset includes information related to demographics, academic performance, mental health factors, and lifestyle attributes. The primary objective of this analysis is to identify patterns and relationships within the dataset and to build predictive models to better understand how these factors contribute to depression among students.

**DATA EXPLORATION:**

***How the Data Was Collected***

The dataset utilized for this analysis was obtained from a publicly accessible online repository, ensuring transparency and reliability. It was originally available in CSV format, a widely accepted and user-friendly data format that facilitates ease of manipulation and analysis. Upon acquisition, the data was carefully examined for completeness and integrity before being imported into a Pandas DataFrame, a versatile data structure in Python commonly used for data analysis. This preprocessing step ensured the data was in a format conducive to exploration, visualization, and statistical modeling. Any missing values, duplicates, or inconsistencies in the dataset were addressed during this stage to enhance its quality and accuracy for subsequent analysis.

***Features Identified for Analysis***

Key features included:

**Gender Distribution**: Countplot revealed a balanced gender distribution.

**Age Distribution**: Histogram indicated age clustering around a mean.

**CGPA by Gender**: Boxplot showcased variance in CGPA across genders.

**Study Satisfaction**: Barplot highlighted correlations between CGPA and satisfaction levels.

**Suicidal Thoughts**: Pie chart showed the percentage of students with suicidal ideations.

**Visualizations**

**Figure 1. Gender Distribution:**

**A close-up of a computer code

Description automatically generated**

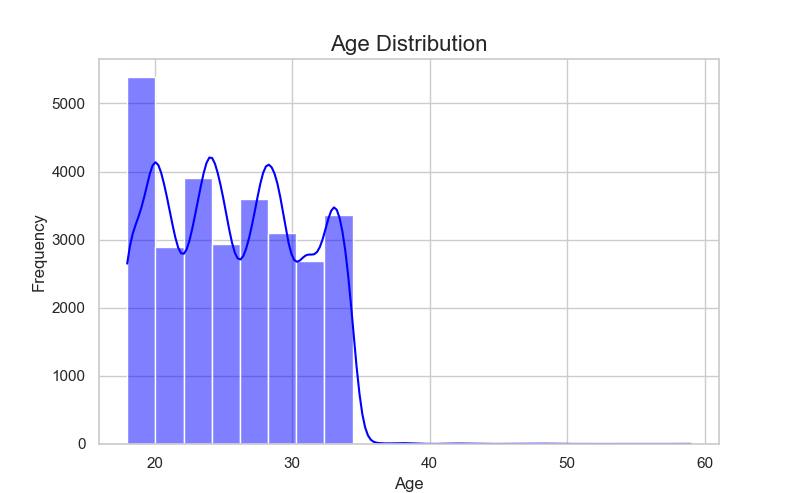


**Visualization**: A bar chart was created to display the distribution of genders in the dataset. The results indicate that the dataset is evenly distributed across genders, which ensures fairness in analysis.

**Insight**: The count plot shows that the dataset includes more males than females. This imbalance could indicate that males might be more likely to participate in surveys related to mental health, potentially skewing the results.

A computer screen shot of a code

Description automatically generated**Figure 2**. **Age Distribution:**



**Visualization:** A histogram was generated to illustrate the age distribution of students. It was observed that the majority of the students fall within the late teenage years to early twenties.

**Insights:** The **Age Distribution** graph shows that the highest frequency of ages is concentrated around the early 20s, particularly ages **18 to 22**. The frequency decreases significantly after **30**, indicating that the dataset primarily includes younger individuals, such as students or early-career professionals.

This trend can be explained by the nature of the dataset, which likely focuses on students. The majority of students fall within the 18–22 age group, which reflects typical college or university ages. The lower frequencies in older age groups suggest fewer older participants in the study.

**Figure 3.** **CGPA Distribution by Gender:**

A close-up of a computer code

Description automatically generated

A chart with blue and orange rectangular boxes

Description automatically generated

**Visualization**: A boxplot was used to analyze the CGPA distribution by gender. It showed no significant difference between male and female students in terms of academic performance.

**Insights:** Female participants exhibit slightly higher CGPA levels compared to males. This difference might be influenced by academic behaviors or pressures specific to gender roles.

**Figure 4.** StudySatisfaction and CGPA by Gender

A close-up of a text

Description automatically generated

A graph of a graph of a person

Description automatically generated with medium confidence

**Visualization**: A barplot was created to explore the relationship between study satisfaction and CGPA, segmented by gender. Students with higher levels of study satisfaction generally exhibited better academic performance.

**Insights**: Despite varying levels of study satisfaction, the CGPA remains high (around 7.5-8) for both genders, indicating that students likely maintain good academic performance irrespective of how satisfied they feel. Students reporting a satisfaction level of 0 tend to have **lower CGPA scores**, especially with higher variability. This suggests that dissatisfaction in studies could have a negative impact on academic performance for a small subset of the group.

**A screen shot of a computer

Description automatically generatedFigure 5.** Suicidal Thoughts Distribution

A blue and red pie chart

Description automatically generated

**Visualization:** pie chart was employed to visualize the proportion of students who reported having suicidal thoughts. The results highlight a concerning proportion of students experiencing mental health issues.

**Insights:** A significant portion of respondents have reported experiencing suicidal thoughts. Specially the women, meaning that the woman is likely thinking about suicide, This alarming statistic suggests the need for immediate mental health interventions and resources for the affected demographic.

**METHODS**

***Pre-Processing Techniques Used***

* **Data Cleaning:** Columns with irrelevant information (e.g., IDs, city) were removed. Missing values were checked and handled.

A black background with white text

Description automatically generated

In the data cleaning phase, unnecessary or irrelevant columns were removed to simplify the dataset. The id column was dropped as it doesn't contribute to the analysis. The City, Degree, Profession, Work Pressure, and Job Satisfaction columns were also removed as they were either not relevant to the study of mental health or needed to be excluded based on the feature selection criteria. By removing these features, we focused the dataset on attributes that directly relate to the students' mental health status and behaviors.

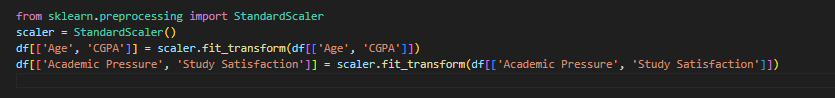
* **Encoding:** Categorical variables, such as gender and dietary habits, were converted into numerical values for modeling.

A computer screen shot of a program code

Description automatically generated

For encoding categorical data, several features were transformed into numerical values to make them suitable for machine learning models. The Gender column was encoded by assigning 0 to Male and 1 to Female. The Sleep Duration and Dietary Habits columns, which contained categorical information, were similarly mapped to numeric values for easier analysis. For the Sleep Duration, categories like Less than 5 hours, 5-6 hours, 7-8 hours, and More than 8 hours were assigned values 0, 1, 2, and 3, respectively. The Dietary Habits were also encoded where Healthy, Unhealthy, and Moderate were mapped to 0, 1, and 3. Additionally, the Have you ever had suicidal thoughts? and Family History of Mental Illness columns were label encoded using the LabelEncoder to convert the categorical answers into binary numeric values, preparing them for machine learning.

* **Scaling:** Features like age and CGPA were standardized using StandardScaler**.**

****

Feature selection involved removing rows and columns that did not contribute meaningfully to the model. Cities with fewer than 400 students were removed to avoid bias from data sparsity, as small sample sizes may lead to unreliable conclusions. Additionally, we filtered the dataset to include only students by keeping rows where the Profession column value was Student. This decision aligns with the objective of studying mental health among students specifically and ensures that the data used for model training is relevant and consistent.

* **Feature Selection:** Cities with fewer than 400 students and non-student professions were excluded.

A screen shot of a computer code

Description automatically generated

Removing Small Cities Excludes cities with fewer than 400 students to maintain consistency and reduce data sparsity.Focusing on StudentsFilters the data to include only Student professions, as the study focuses on mental health among students.

**CONCLUSION**

This analysis underscores the critical factors associated with student depression. The Random Forest model developed during this analysis achieved high accuracy, proving to be an effective tool for identifying students at risk of depression. Among the most influential factors were a family history of mental illness, levels of academic pressure, and study satisfaction, all of which play a crucial role in determining a student’s mental health.

**REFERENCE**

Hopesb. (2022). *Student Depression Dataset* [Data set]. Kaggle. Retrieved from <https://www.kaggle.com/datasets/hopesb/student-depression-dataset/data>

Mia, A. (2023). *Student Depression Dataset Analysis & Hybrid Model* [Code notebook]. Kaggle. Retrieved from <https://www.kaggle.com/code/arifmia/student-depression-dataset-analysis-hybrid-model>