Predicting Student Depression: A Collaborative Data Analysis Report

Main Objective: Develop a predictive model to determine students at risk of depression by leveraging demographic, social, financial, and academic indicators

Business Objectives	Analytical Objectives
- Who are the main stakeholders who will benefit most from our findings? - Which social, demographic, and academic variables should be considered for predicting students at risk of depression? - Which groups of students can be identified as being at risk of depression? - Which groups of students would benefit from additional mental health support? - What intervention strategies can be developed, and which institutions or groups of people can implement these strategies? - What strategies can be proposed to help students better manage their mental health?	- Which columns of data are complete, consistent, and reasonable? - Describe the distribution of depression amongst the student population? Is it a normal distribution? - How strongly are the selected variables correlating with the depression score? - How are the selected variables correlating with another? - What is the best performing ML model for predicting students at risk of depression? - Is the best performing ML model based solely on demographic factors? Social factors? Academic factors? Or a combination of different factors? - What is the margin of error in these predictions?

Data

The data was collected by surveying anonymous students across various educational institutions in India. Those who were surveyed provided information on their mental health, academics, lifestyle habits, and demographics.

Key Highlights of Dataset

- *Dataset:* There were 27901 records present. It was mostly complete, with only 3 missing values in the "Financial Stress" column.
- *Depression (Target Variable):* This binary column shows if a student is experiencing depression ("0" = No and "1" = Yes).
- *Other Binary Variable:* The "Have you ever had suicidal thoughts?" & "Family History of Mental Illness" consist of "Yes" and "No" string values.
- *Dataset Source:* Student Depression Dataset (https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset)

Figure: Summary Table of Variables

Categories	Variables	
	Categorical	Numerical
Descriptive Fields	ID	
Demographic	Gender City	Age
Academic	Degree	Academic Pressure CGPA Study Satisfaction
Financial		Financial Stress Work Pressure Job Satisfaction
Social	Sleep Duration Dietary Habits Have you ever had suicidal thoughts ? Family History of Mental Illness	Work/Study Hours Depression **(Target Variable)**

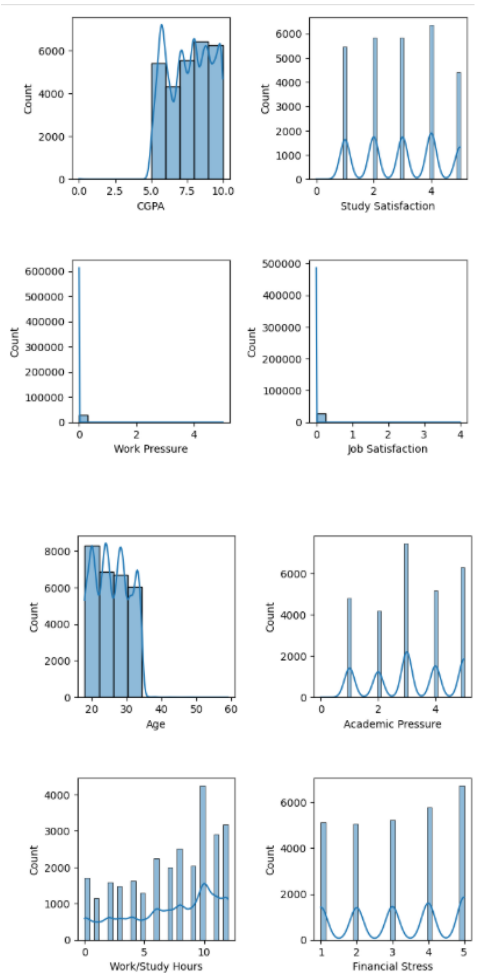
Exploratory Data Analysis

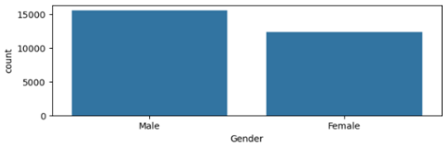
Upon reviewing the data types, the "Financial Stress" variable was incorrectly labeled as an "object" type but should be a "float" type. The categorical variables "Have you ever had suicidal thoughts?" and "Family History of Mental Illness" are binary while "Sleep Duration" and "Dietary Habits" are ordinal. All can be converted into numerical data (see the Data cleaning section for details on addressing errors, missing values, and encoding). Moreover, "Financial Stress" had 3 missing values but all other variables in the dataset were complete with no missing values found. The results of the "Depression" target variable were observed to be balanced with no heavy skewing towards one outcome and no duplicate entries. All changes to the dataset were documented. The numerical data was not normally distributed. As such, a logistical regression model was selected, which doesn't assume variables are normally distributed.

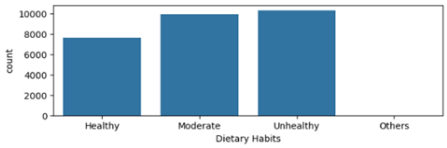
Validating Categorical and Numerical Variables

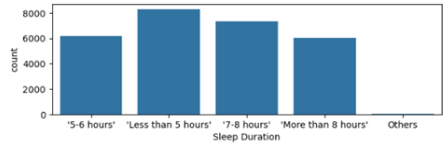
Regarding the numerical variables, "Study Satisfaction", "Financial Stress", and "Academic Pressure" have evenly spread data. "Study Satisfaction" looks normally distributed, while "Academic Pressure" has a bimodal distribution with the 2 peaks being non-symmetrical. "CGPA" and "Work/Study Hours" are heavily and moderately left-skewing, respectively. "Age" is heavily right-skewed. Finally, "Work Pressure" and "Job Satisfaction" are heavily right-skewed, with many data points near zero.

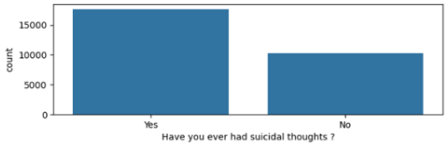
Moreover, there is a good balance of answers to the categorical questions. The exception was the "profession" variable where the response was "student", which was expected. Furthermore, "Class of 12" was a strongly dominant response given in the "degree" data column. After analyzing the bar graphs, it was found most of the respondents were male, had unhealthy dietary habits, slept less than 5 hours, and had suicidal thoughts, among other items.

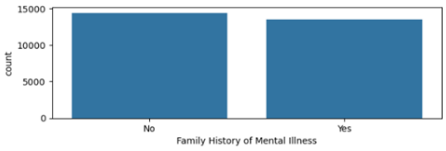


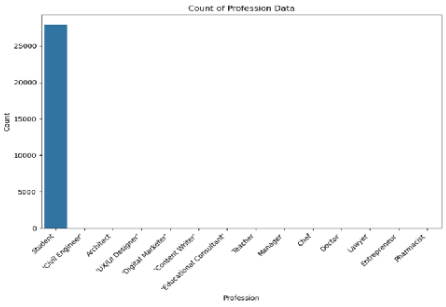


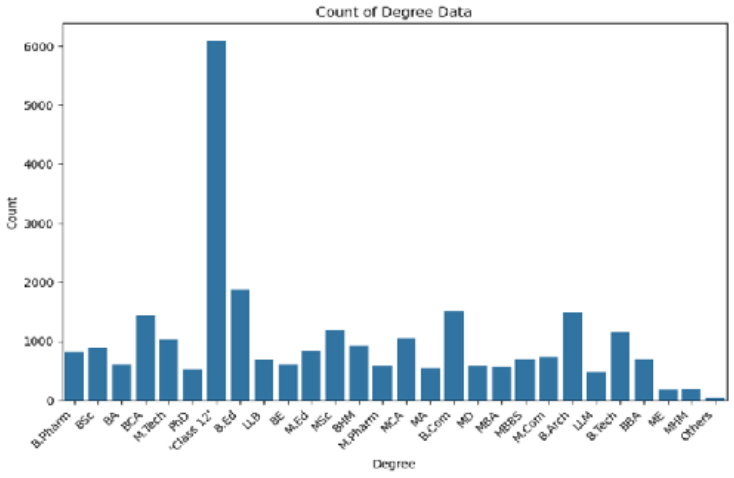




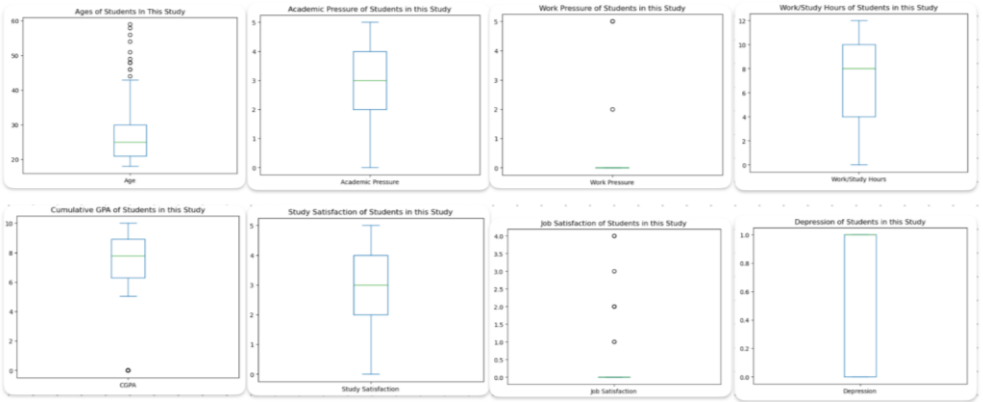








Outlier Analysis: Box Plots for Numerical Variables

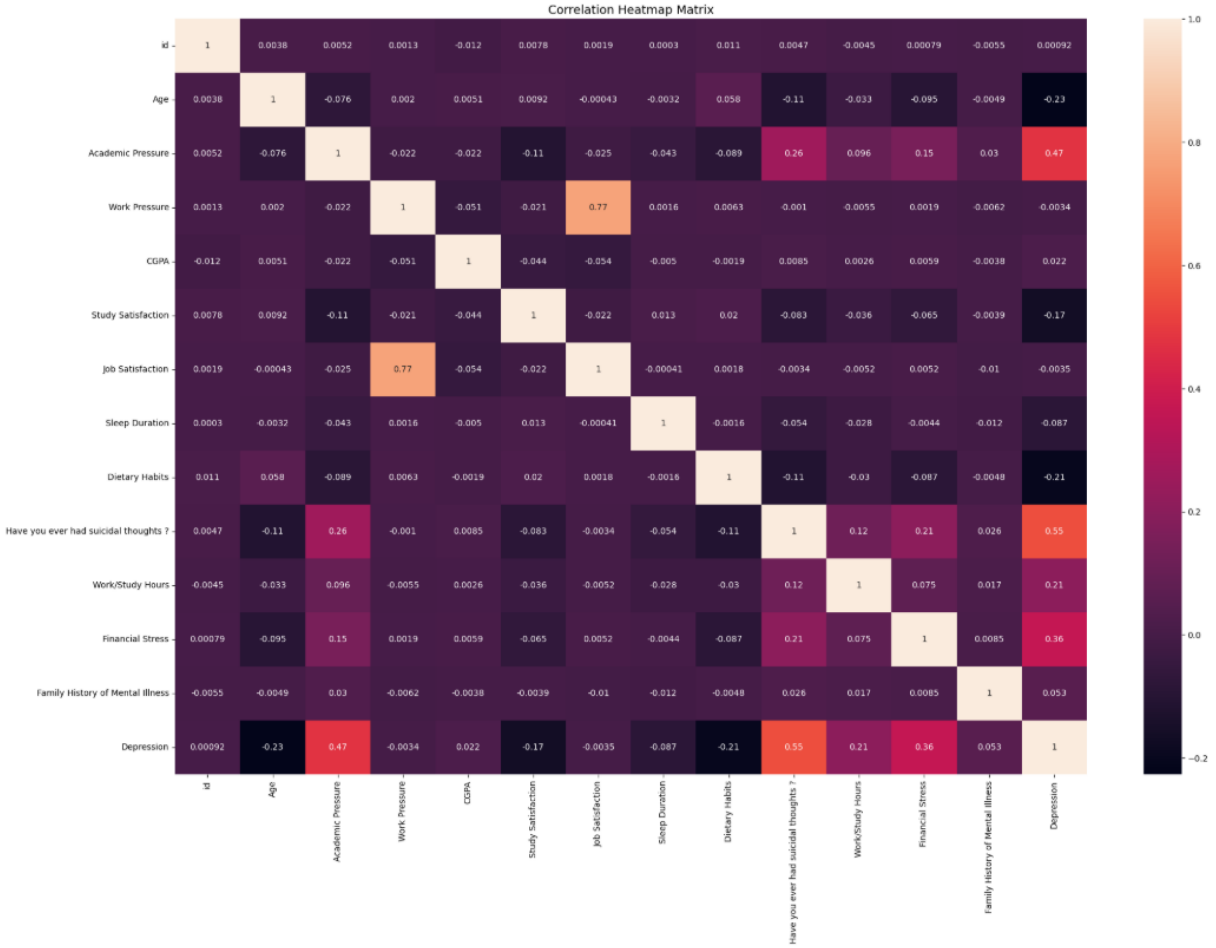


After reviewing the boxplots above, there are several variables (Depression, Work/Study Hours, Study Satisfaction, and Academic Pressure) that do not have outliers. Depression does not have any outliers because it is binary in its collection through this study, and with the other variables it is likely the data is closely related. While Job Satisfaction, Work Pressure, CGPA, and Age have outliers present and this could indicate there are students in this study who feel they have great job satisfaction or work pressure, a very small number of students are academically underachieving, and with age it appears there are a couple of mature students. None of these are concerning and are reflective of a normal post-secondary school environment; outliers in the numerical dataset can be kept, given they can be plausibly explained.

Correlation Analysis

A comprehensive heatmap showing all original and new numerical data was generated. The "Depression" target variable was revealed to have the strongest relationship with "Have you ever had suicidal thought?", "Academic Pressure", "Financial Stress", "Work/Study Hours", "Dietary Habits", and "Study Satisfaction". These independent variables have a moderate - weak positive correlation with depression except for "Dietary Habits" and "Study Satisfaction", which is a negative correlation.

Features	Correlation Coefficient
Have you ever had suicidal thought?	0.55
Academic Pressure	0.47
Financial Stress	0.36
Work/Study Hours	0.21
Age	-0.23
Dietary Habits	-0.21
Study Satisfaction	-0.17



Data Cleaning and Processing

As mentioned in the exploratory data analysis section, 3 missing values were identified for "Financial Stress". These were addressed by inputting median values in place of the missing values. This ensured the estimated values weren't significantly affected by potential outliers.

The categorical features, "Have you ever had suicidal thoughts?" and "Family History of Mental Illness" are binary features with "Yes" & "No" values. To enable the developed model to learn from this data, these values were encoded into numerical data similar to the "Depression" target variable where No = "0" and Yes = "1".

Moreover, the "Sleep Duration" & "Dietary Habits" categorical features can be assumed as ordinal and converted into numerical data. For example, "Unhealthy" is of a lower order than "Moderate" in "Dietary Habits" while "7-8 hours" is of a higher order than "5-6 hours" in "Sleep Duration". After the conversion, 18 and 12 missing values were identified for "Sleep Duration" & "Dietary Habits", respectively. Similar to "Financial Stress", these missing values were addressed by imputing median values.

Sleep Duration	
Categorical	***Numerical***
Less than 5 hours	0
5-6 hours	1
7-8 hours	2
More than 8 hours	3
Others	NaN (missing values)

Dietary Habits	
Categorical	**Numerical**
Unhealthy	0
Moderate	1
Healthy	2
Others	NaN (missing values)

Final Logistic Regression Model

We chose to use a logistic regression model because logistic regression models are built to predict probability-based outcomes and are highly applicable when predicting a binary outcome such as whether a student is depressed and they are less prone to overfitting compared to other machine learning models. It allowed us to see based on the selected features listed below whether we could accurately predict if a student was or was not depressed even with a larger dataset, and it allowed us to see if the features we chose were relevant to predicting depression.

Selected Features

Based on the correlation matrix, numerical variables with a low correlation to "Depression" were removed: "Family History of Mental Illness", "Sleep Depression", "id", "Work Pressure", "CGPA", "Job Satisfaction", "Sleep Duration". The "Age" feature was also dropped because depression can occur at any life stage. Lastly, "City", "Gender", "Degree", and "Profession" were not considered due to being categorical variables.

Selected Features
Have you ever had suicidal thought?
Academic Pressure
Financial Stress
Work/Study Hours
Dietary Habits
Study Satisfaction

Model Validation

We validated the logistic model using a train-test split of 70/30, holding 30% of our data for testing purposes. This resulted in a 84% accuracy score, meaning high confidence in our model. We are satisfied with this accuracy score, as accuracy scores that are too high indicate overfitting which we aim to avoid. We also validated this model using the k-folds cross validation method which gave us an accuracy score of 83.69%.

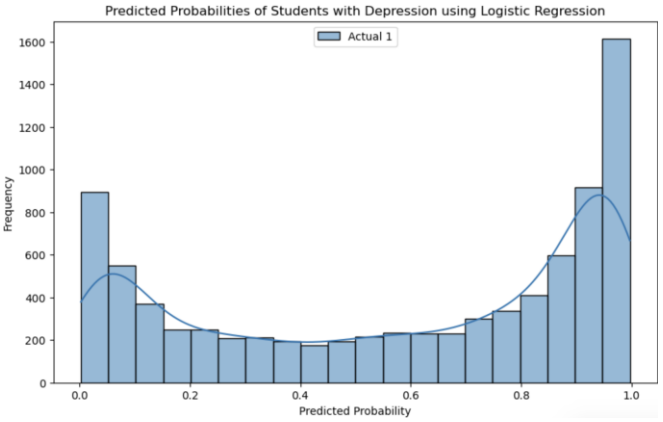
We also validated the random forest model using the k-folds cross validation method which gave us an accuracy score of 80.55%. We chose this method as it maximizes/ensures all data from the original data set is being used for both training and test sets, as opposed to the train test split method which requires us to save a certain portion of the data for testing. Using the k-folds cross validation method allows us to not make that tradeoff, providing a more accurate estimate for model quality and performance, and reducing overfitting. We created a random forest model as it was a model that was discussed to be potentially well suited for the classification problem that we sought to solve and wanted to explore how it compared to our logistic regression model.

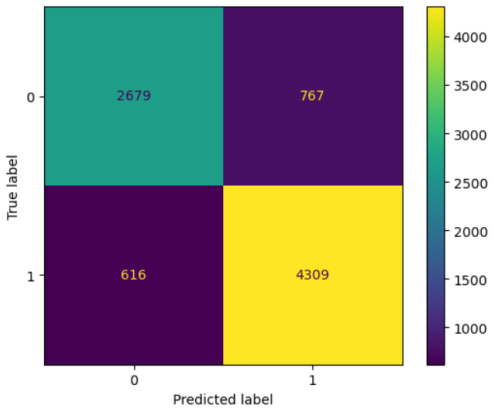
As the accuracy score when conducting the k-folds cross validation method for both methods resulted in a higher percentage of 83.69% in the logistic model vs 80.55% in the random forest model, this led to us choosing the logistic regression model as our final choice of model. This combination of the logistic regression model, when validating using the k-folds cross validation method, was more suited for our data, the target outcome, and our main objective, especially since we are looking at a classification problem.

Model Prediction

The logistic regression model for predicting student depression performed well, hitting an overall accuracy of 84%. The classification report shows it's particularly good at identifying depressed students, with a precision of 85%, meaning it correctly classified 85% of the students it predicted as depressed. The table showed a recall of 87%, which identified 87% of all truly depressed students. This balance gives it a solid F1 score of 0.86, reflecting the model's ability to maintain both high precision and recall, which is critical for minimizing both false positives and false negatives.

For non-depressed students, the model had a precision of 81% and a recall of 78%, resulting in an F1-score of 0.79. This means it can occasionally misclassify some students, likely due to overlapping variables or errors in the data, but overall, it's clearly better at catching depressed students than missing them. The overall metrics confirm that the model is particularly strong at identifying at-risk students, making it a valuable tool for early intervention and targeted mental health support.





Features	Model Coefficients
Have you ever had suicidal thought ?	1.202812
Academic pressure	1.141133
Financial Stress	0.814287
Work/Study Hours	0.429108
Study Satisfaction	-0.320659
Dietary Habits	-0.436009

```
Classification Report:
                precision    recall  f1-score   support

Not Depressed       0.82      0.79      0.80      3500
    Depressed       0.85      0.88      0.86      4871

     accuracy                           0.84      8371
    macro avg       0.84      0.83      0.83      8371
 weighted avg       0.84      0.84      0.84      8371
```

Conclusion

In conclusion, our logistic regression model for student depression performed exceptionally well, achieving an impressive 83.69% accuracy. This high accuracy makes it a strong tool for identifying at-risk students. When selecting features, we deliberately chose to set aside certain variables, like age and gender, as they showed weaker correlations with depression and didn't align with the core psychological and social factors we wanted to emphasize. Instead, we focused on more meaningful variables that were highly correlated with depression, like academic pressure, financial stress, and work/study hours. Our logistic regression model, also outperformed the Random Forest model, highlighting its effectiveness as the preferred choice for this task. We also acknowledge that the topic of student depression is heavily nuanced. Our data was pulled from one country and thus results can vary drastically given sociopolitical, cultural, and economic factors. As the topic speaks to many of us, it was extremely interesting to delve into, but we also recognize the limitations of our dataset in comparison to the depth and extensiveness of the topic and research of student depression.

Next Steps

Looking ahead, our recommendation for educational institutions is to use these insights to craft more targeted support programs, enhance early intervention strategies, and allocate mental health resources more effectively for students. This will ensure a comprehensive approach to supporting student well-being.