Predicting Depression Risk Among Students Using Machine Learning

Presented By : Bryan Mejia, Roger Leung, Ahmed Ali, Abul Hasan

Project Motivation

- Background: Depression often goes unnoticed. Many students push through silently, believing their struggles are just part of the college experience. The normalization of stress and burnout can lead to worsen mental health. With the rise of data science, we have the tools to model and analyze and better understand the multifaceted issues through structured data obtain through student surveys.
- Problem: Through our efforts of EDA and modeling is it possible to outline all the key factors that seem to have a connection to the development of depression?
- Goals: Determining if there are key contributors to symptoms of depression during a student's academic career. And in helping to find a deeper understanding of the pivot points of depression in order improve the mental health of students so that can lead improve of a student's well being across the nation.
- Result: This project will result in an interpretable analysis of key contributors to depression and models to help predict if a student may have depression.

Outline

- Introduction of our dataset
- Exploration and visualizations
- Feature selection and engineering
- Machine learning models
- Model Evaluation

Approach

- Our approach to managing and analyzing this dataset will involve a systematic EDA using Python within a Jupyter Notebook environment. The initial phase will involve data cleaning and preprocessing to handle any missing values, outliers and inconsistencies. We will then perform descriptive statistics to summarize the key characteristics of each feature and identify potential distributions and relationships.
- For our statistical analysis, we plan to utilize a combination of visualization techniques and inferential statistics. We will generate various plots such as histograms, box plots, scatter plots and correlation matrices to represent data distributions, identify correlations between variables and highlight potential patterns.

Introduction to our dataset

- The data that will be used is a dataset taken from OpenML labeled “student_depression_dataset” (OpenML) authored by Israel Campero Jurado. This dataset contains self-reported information and demographic details, including but not limited to: Age, City, Profession, Academic Pressure, Work Pressure, Study Satisfaction and many more. The features provided within this dataset consist of numerical and categorical variables that we will need to further analyze to support our findings.

Statistics of our data set

- Depression data set contains

  27,900 records and 18 features.

```
Shape of the dataset: (27900, 18)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27900 entries, 0 to 27899
Data columns (total 18 columns):
 #   Column                                 Non-Null Count  Dtype
---  ------                                 --------------  -----
 0   id                                     27900 non-null  int64
 1   Gender                                 27900 non-null  object
 2   Age                                    27900 non-null  float64
 3   City                                   27900 non-null  object
 4   Profession                             27900 non-null  object
 5   Academic Pressure                      27900 non-null  float64
 6   Work Pressure                          27900 non-null  float64
 7   CGPA                                   27900 non-null  float64
 8   Study Satisfaction                     27900 non-null  float64
 9   Job Satisfaction                       27900 non-null  float64
 10  Sleep Duration                         27900 non-null  object
 11  Dietary Habits                         27900 non-null  object
 12  Degree                                 27900 non-null  object
 13  Have you ever had suicidal thoughts ?  27900 non-null  object
 14  Work/Study Hours                       27900 non-null  float64
 15  Financial Stress                       27900 non-null  object
 16  Family History of Mental Illness       27900 non-null  object
 17  Depression                             27900 non-null  int64
dtypes: float64(7), int64(2), object(9)
```

Statistics of our data set cont.

Sample of the data set:

	id	Gender	Age	City	Profession	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/Study Hours	Financial Stress	Family History of Mental Illness	Depression
0	8	Female	24.0	Bangalore	Student	2.0	0.0	5.90	5.0	0.0	'5-6 hours'	Moderate	BSc	No	3.0	2.0	Yes	0
1	26	Male	31.0	Srinagar	Student	3.0	0.0	7.03	5.0	0.0	'Less than 5 hours'	Healthy	BA	No	9.0	1.0	Yes	0
2	30	Female	28.0	Varanasi	Student	3.0	0.0	5.59	2.0	0.0	'7-8 hours'	Moderate	BCA	Yes	4.0	5.0	Yes	1
3	32	Female	25.0	Jaipur	Student	4.0	0.0	8.13	3.0	0.0	'5-6 hours'	Moderate	M.Tech	Yes	1.0	1.0	No	0
4	33	Male	29.0	Pune	Student	2.0	0.0	5.70	3.0	0.0	'Less than 5 hours'	Healthy	PhD	No	4.0	1.0	No	0
5	52	Male	30.0	Thane	Student	3.0	0.0	9.54	4.0	0.0	'7-8 hours'	Healthy	BSc	No	1.0	2.0	No	0
6	56	Female	30.0	Chennai	Student	2.0	0.0	8.04	4.0	0.0	'Less than 5 hours'	Unhealthy	'Class 12'	No	0.0	1.0	Yes	0
7	59	Male	28.0	Nagpur	Student	3.0	0.0	9.79	1.0	0.0	'7-8 hours'	Moderate	B.Ed	Yes	12.0	3.0	No	1
8	62	Male	31.0	Nashik	Student	2.0	0.0	8.38	3.0	0.0	'Less than 5 hours'	Moderate	LLB	Yes	2.0	5.0	No	1
9	83	Male	24.0	Nagpur	Student	3.0	0.0	6.10	3.0	0.0	'5-6 hours'	Moderate	'Class 12'	Yes	11.0	1.0	Yes	1

Statistics of our data set cont.

- Features in our data set

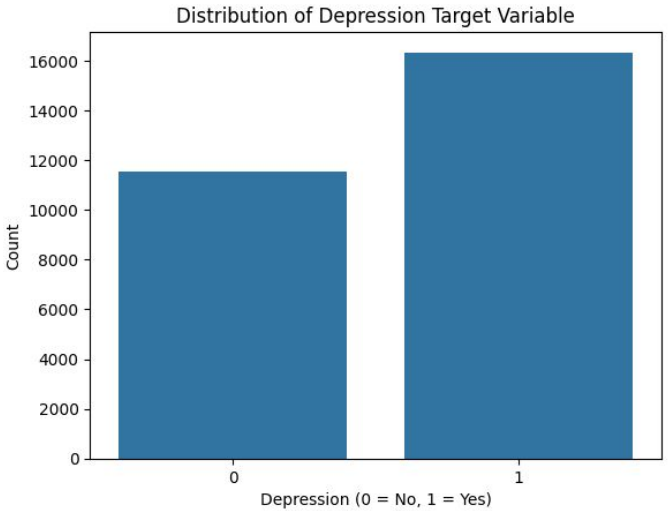
Feature Name	Type	Unique Values	Missing Values	Description
Depression (target)	Numeric	2	0	Binary target variable indicating presence (1) or absence (0) of depression.
id	Numeric	27901	0	Unique identifier for each individual; not useful for modeling.
Gender	String	2	0	Gender of the respondent. Categorical feature.
Age	Numeric	34	0	Age in years.
City	String	52	0	City of residence.
Profession	String	14	0	Current professional role (e.g., student, teacher). Categorical.
Academic Pressure	Numeric	6	0	Self-reported academic pressure (likely ordinal).
Work Pressure	Numeric	3	0	Self-reported work pressure level (likely ordinal).
CGPA	Numeric	332	0	Academic performance score. Continuous.
Study Satisfaction	Numeric	6	0	Satisfaction level with study experience (ordinal).
Job Satisfaction	Numeric	5	0	Self-reported satisfaction with job or work (ordinal).
Sleep Duration	String	5	0	Reported average sleep duration. Consider converting to numeric hours.
Dietary Habits	String	4	0	Type or regularity of dietary habits. (Categorical)
Degree	String	28	0	Level of education (ordinal)

Exploratory Data Analysis (EDA)

- Through univariate analysis, we explore how each feature is distributed independently. Most of our features are categorical or ordinal (e.g., levels of stress, presence/absence of family history, or emotional health), we use bar plots to visualize the count of each category. This helps identify unusual distributions in individual variables before modeling.

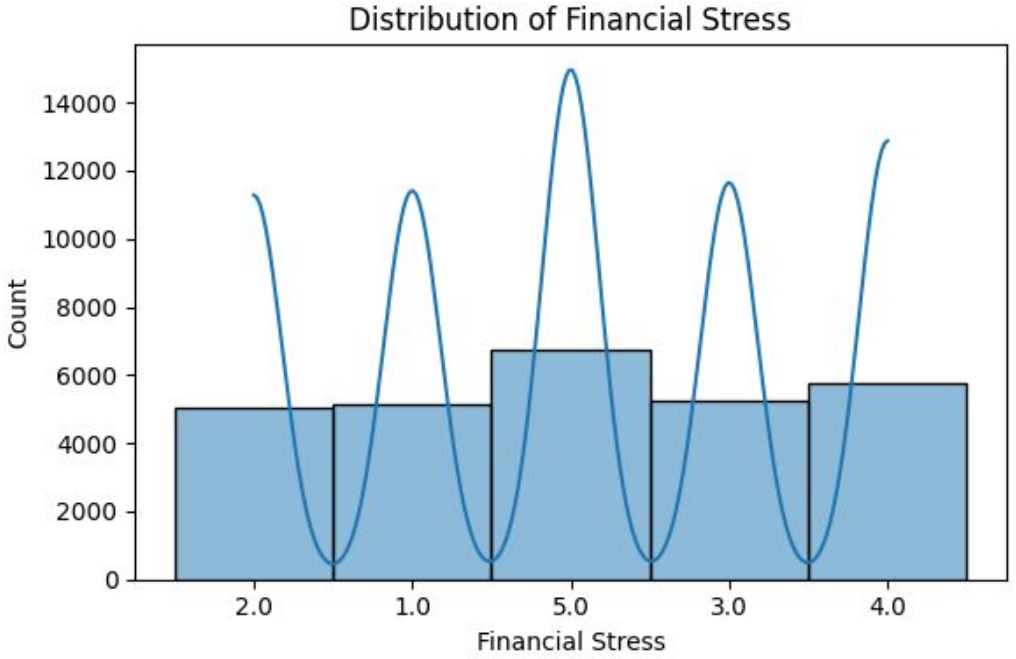
EDA cont.

- Depression (Target Variable): We observe that a greater number of students report with depressive symptoms, with a smaller proportion reporting of no. This imbalance suggests that we may need to consider it during modeling, especially when using performance metrics like Recall and F1-Score.



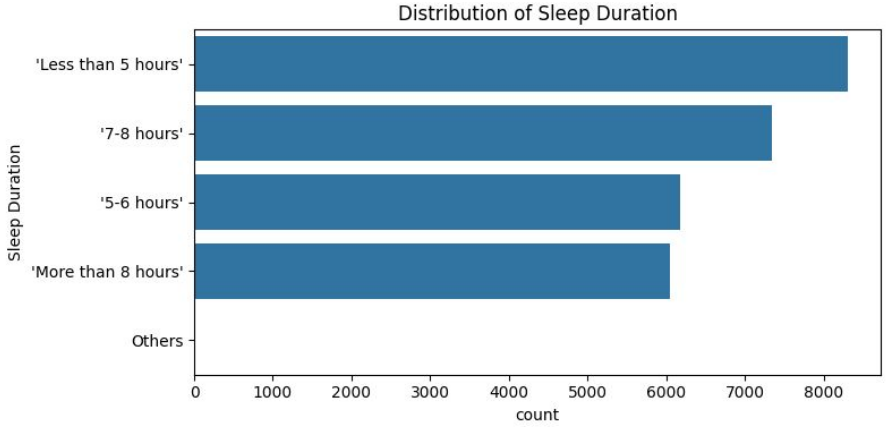
EDA cont.

Financial Stress: The distribution shows a noticeable number of students experiencing moderate to high financial stress, which is a known contributor to mental health issues. This variable appears to be well distributed across categories, making it a strong candidate for predictive analysis.



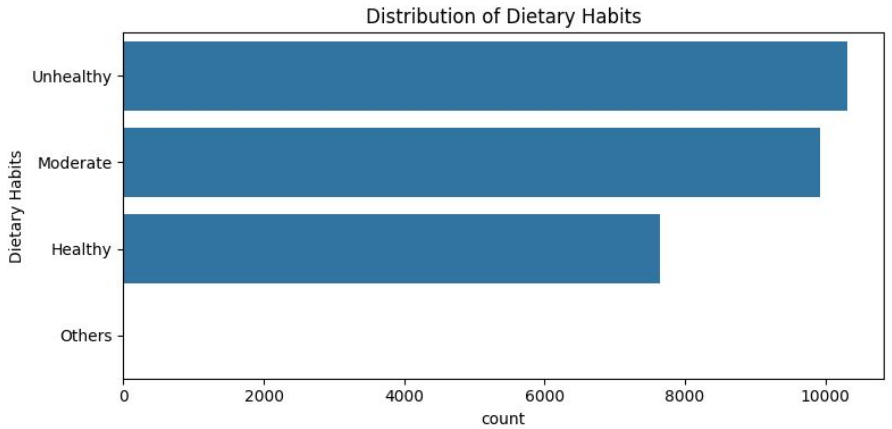
EDA cont.

Sleep Quality: There is a wide spread in reported sleep quality. However, many students report poor or average sleep, which correlates with known triggers for mental fatigue and depressive symptoms. This variable will be particularly important in our model.



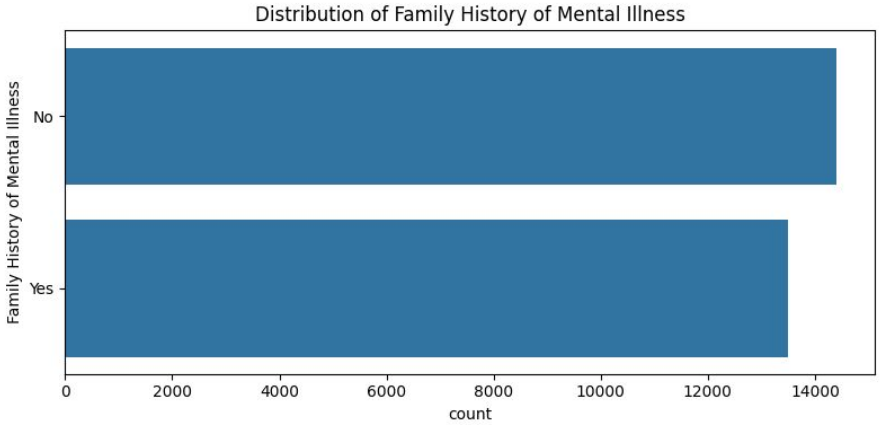
EDA cont.

- Dietary habits: There is a larger portion of students reporting unhealthy diets, which could indirectly contribute to poor emotional well-being.



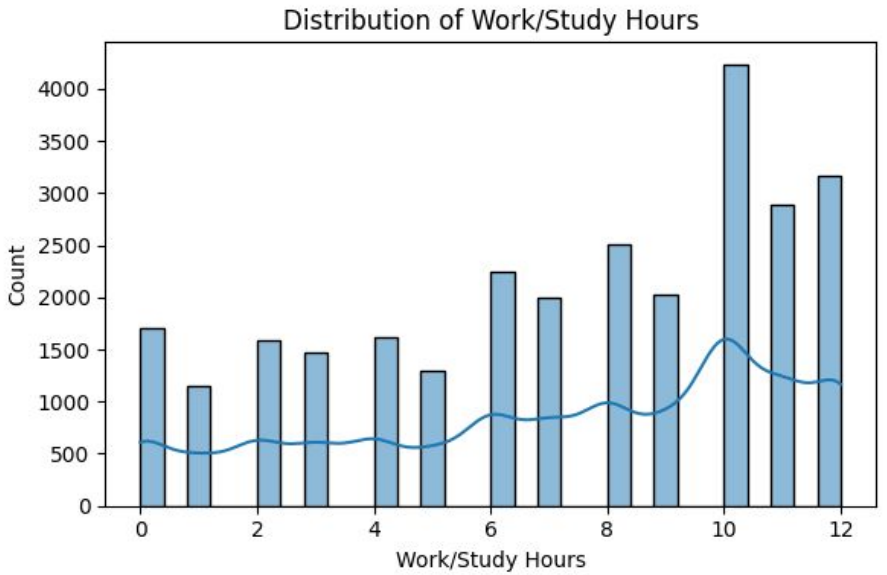
EDA cont.

Family History of Mental Illness: A large number of students report no family history, but there is a significant group that does — showing potential influence from genetics or environment. This will be tested for statistical association with depression using chi-square analysis.



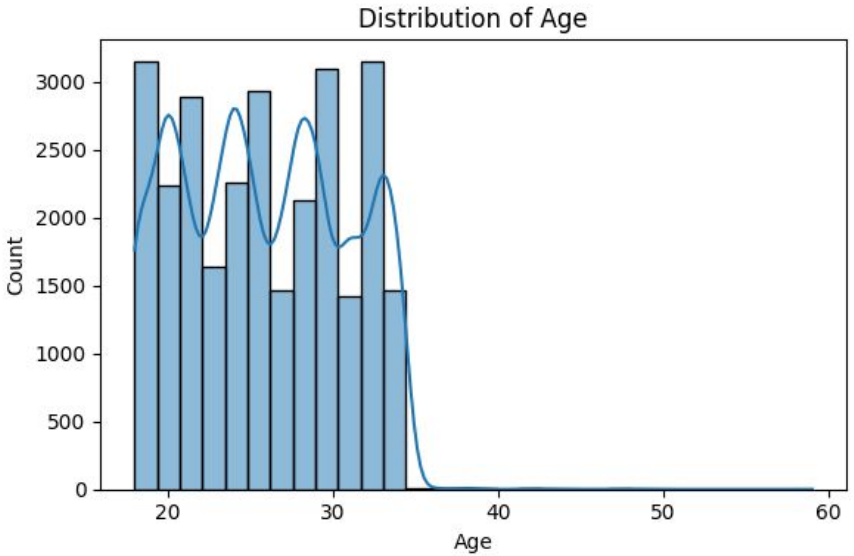
EDA cont.

- The distribution of work/ study skews left. This shows on average students on working/ studying longer hours per day.



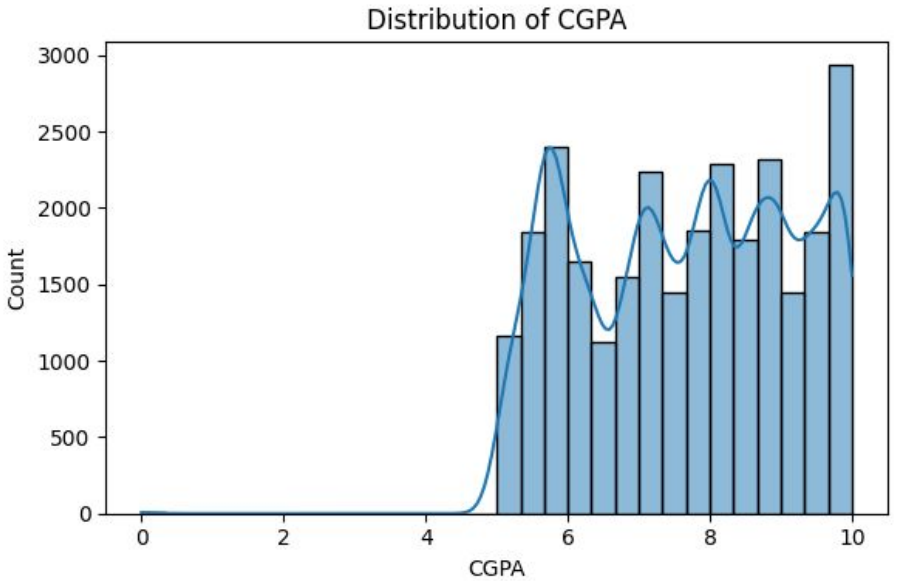
EDA cont.

- The demographic of this data hovers between late teens to mid thirty adults.



EDA cont.

- The cumulative GPA is distributed between “good” and “excellent”. This can tell us that students are studying longer hours to achieve better grades at the cost of sleep and mental health.



Feature Selection and Engineering

- Converting missing variables to nan
- Choosing median to imputethe missing values in financial stress feature.
- We did not implement specific outlier detection and removal techniques. This decision was made to retain as much of the original data as possible, as potential "outliers" in the context of factors influencing depression (such as very high academic pressure or work hours).

Feature Selection and Engineering cont.

- Correlation matrix

```
=== TOP CORRELATIONS WITH DEPRESSION =
Academic Pressure      0.474816
Financial Stress       0.363655
Age                   -0.226478
Work/Study Hours       0.208604
Study Satisfaction    -0.167954
CGPA                   0.022184
Job Satisfaction      -0.003481
Work Pressure         -0.003350
Name: Depression, dtype: float64
=== EDA KEY FINDINGS ===
Dataset shape: (27900, 18)
Depression rate: 58.5%
```

Feature Selection and Engineering cont.

- Using Chi square (greater the number suggest rejecting the null hypothesis)
- Using p-values (closer to 0 suggests rejecting the null hypothesis)
- Features to consider: Sleep duration, Degree, Suicidal thoughts, and Family history

```
Gender: Chi2 = 0.08, p = 0.7772
City: Chi2 = 187.99, p = 0.0000
Profession: Chi2 = 14.29, p = 0.3538
Sleep Duration: Chi2 = 276.90, p = 0.0000
Dietary Habits: Chi2 = 1203.15, p = 0.0000
Degree: Chi2 = 531.56, p = 0.0000
Have you ever had suicidal thoughts ?: Chi2 = 8323.25, p = 0.0000
Family History of Mental Illness: Chi2 = 79.52, p = 0.0000
```

Feature Selection and Engineering cont.

- Average Work/Study Hours and Financial Stress by Sleep Duration
- We can observe if there are notable differences in these averages across different sleep durations. For instance, we might see if individuals reporting less sleep tend to have higher work/study hours or financial stress, which could be contributing factors to mental health issues. Conversely, those with longer sleep durations might show lower averages in these areas. These insights can inform our understanding of the interplay between lifestyle factors and potential depression risk.
-

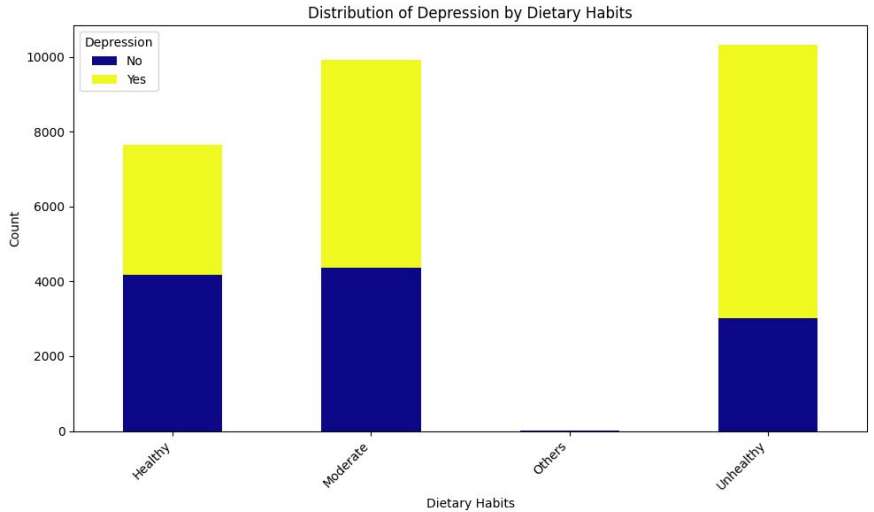
Average Work/Study Hours and Financial Stress by Sleep Duration:

Sleep Duration	Work/Study Hours	Financial Stress
'5-6 hours'	7.282918	3.113070
'7-8 hours'	7.266131	3.172611
'Less than 5 hours'	7.194705	3.152226
'More than 8 hours'	6.845467	3.112177
Others	6.777778	2.666667

Feature Selection and Engineering cont.

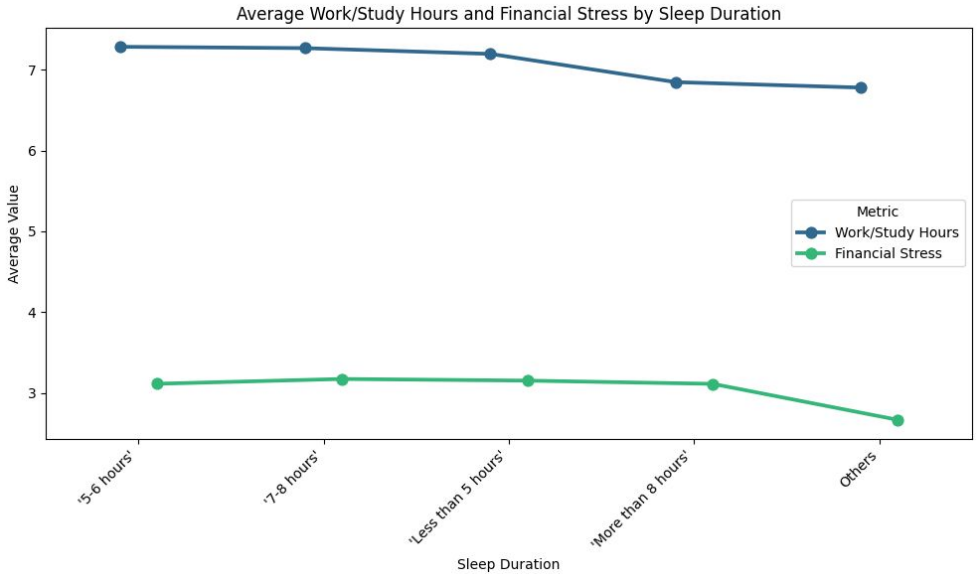
- Distribution of depression by dietary habits
- There is a big difference in unhealthy diet category.

Depression	0	1
Dietary Habits		
Healthy	4178	3472
Moderate	4363	5558
Others	4	8
Unhealthy	3020	7297



Feature Selection and Engineering cont.

- Average Work/Study Hours and Financial Stress by Sleep Duration
- we can see if there are discernible trends, such as whether individuals reporting shorter sleep durations tend to have higher average work/study hours or financial stress.



K-Nearest Neighbors(KNN)

- KNN can be suitable for this classification problem because it doesn't make assumptions about the underlying data distribution, which can be beneficial if the relationships between features and the target variable are non-linear or complex. However, its performance can be sensitive to the choice of 'k' and the distance metric used, and it can become computationally expensive with large datasets, especially during the prediction phase.

K-Nearest Neighbors Classification Report:

	precision	recall	f1-score	support
0	0.40	0.34	0.36	2261
1	0.59	0.65	0.62	3319
accuracy			0.52	5580
macro avg	0.49	0.49	0.49	5580
weighted avg	0.51	0.52	0.52	5580

Decision Tree Classifier

- Decision Trees are relatively easy to understand and interpret, as they can be visualized as a tree structure. They can capture non-linear relationships and interactions between features. However, they can be prone to overfitting, especially with complex trees, and small variations in the data can lead to significantly different tree structures. We will evaluate its performance to see how it compares to the other models.

Decision Tree Classification Report:

	precision	recall	f1-score	support
0	0.71	0.72	0.72	2261
1	0.81	0.80	0.81	3319
accuracy			0.77	5580
macro avg	0.76	0.76	0.76	5580
weighted avg	0.77	0.77	0.77	5580

Random Forest Classifier

- Before training some machine learning models, particularly those that are sensitive to the scale and distribution of features, it's crucial to scale our data.

```
=== Random Forest Classifier ===

Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.76      0.79      2261
           1       0.84      0.89      0.86      3319

    accuracy                           0.83      5580
   macro avg       0.83      0.82      0.83      5580
weighted avg       0.83      0.83      0.83      5580

Random Forest Balanced Accuracy: 0.8222
```

Support Vector Machine Classifier

- SVM finds the hyperplane in an N-dimensional space (where N is the number of features) that distinctly classifies the data points.
- We will train the model on the scaled data and evaluate its performance, paying attention to how it handles the potential data imbalance using appropriate metrics.

```
=== Support Vector Machine (SVM) Classifier ===

SVM Classification Report:
              precision    recall  f1-score   support

           0       0.57      0.67      0.62      2261
           1       0.74      0.66      0.70      3319

    accuracy                           0.66      5580
   macro avg       0.66      0.66      0.66      5580
weighted avg       0.67      0.66      0.67      5580

SVM Balanced Accuracy: 0.6629

SVM Cross-Validation (Balanced Accuracy) Scores (3-fold):
[0.38490694 0.56987401 0.67479675]
Mean CV Balanced Accuracy: 0.5432
```

Naive Bayes Classifier

- We used the GaussianNB variant of Naive Bayes.
- Naive Bayes model is struggling significantly, particularly with correctly identifying the positive class (Depression=1). The low precision, recall, and f1-score for the 'Depression' class, along with a low balanced accuracy, suggest that the simplifying assumption of feature independence might not hold true for this dataset, leading to poor performance.

```
=== Naive Bayes Classifier ===

Naive Bayes Classification Report:
              precision    recall  f1-score   support

           0       0.40      0.99      0.58      2261
           1       0.46      0.00      0.01      3319

    accuracy                           0.40      5580
   macro avg       0.43      0.50      0.29      5580
weighted avg       0.44      0.40      0.24      5580

Naive Bayes Balanced Accuracy: 0.4988
```

Models Summary

	Model	Accuracy	Balanced Accuracy (Test)	Mean CV Balanced Accuracy
2	Decision Tree	0.96	0.9600	N/A (Not explicitly calculated)
0	Logistic Regression	0.84	0.8300	N/A (Not explicitly calculated)
3	Random Forest	0.83	0.8222	0.8305
1	K-Nearest Neighbors	0.82	0.8100	N/A (Not explicitly calculated)
4	SVM (Linear Kernel)	0.66	0.6629	0.5432
5	Naive Bayes	0.40	0.4988	N/A (Not explicitly calculated)

Conclusion

- The use of balanced_accuracy and class_weight='balanced' was crucial for evaluating and training models on this potentially imbalanced dataset, providing a more reliable performance assessment than standard accuracy alone.
- While the Decision Tree showed very high performance on the test set, its potential for overfitting should be considered, especially if cross-validation results were significantly lower than test performance. Further hyperparameter tuning and cross-validation with appropriate metrics would be a valuable next step for all models to confirm their generalization ability.

Behind the Papers: A Student's Journey Navigating the Hidden Struggles of College Mental Health

At a bustling Queens College filled with deadlines, grades, and late-night study sessions, mental health is often an invisible battle. Students submit papers with a signs of relief, while juggling part-time jobs. But behind those signs of relief, there are many of those who are quietly suffering.

Mental health challenges, especially depression often go unnoticed and unspoken. Many students push through silently, believing their struggles are just part of the college experience.

But recent studies show that depression is more common among students than many realize, and most who suffer don't receive the support they need. It is clear that that more research in understanding, identifying, and educating students and parents on depression is crucial. For some people, the idea of using data or research to address such a personal, emotional issue can feel cold or disconnected. After all, depression is not just a statistic, it is very real thing.

Now, more institutions are now collecting mental health data on sleep, study habits, social support, and emotional wellbeing. Patterns are emerging, and with them, insights that can drive some meaningful change. These changes can lead to real outcomes, such as earlier detection, less stigma, and a greater sense of community. Students can begin to realize they are not alone and that help is both available and acceptable to receive help. And through that, we can give students a better chance to succeed academically and to live well.

Timeline & Management

Search for a Dataset: 6/21/2025

Proposal: 6/22/2025

Deliverable: 6/24/2025

Q & A

Will have time to answer any questions about our presentation