CS-513
Student Depression

Andrew Krasinski
Marcus Hom
Uros Nikolic

Uros Nikolic
20017063

Andrew Krasinski
10475994

Marcus Hom
20014292

01

Background

Problem Statement

In modern times, students are under a lot of stress, and many students are diagnosed with depression. We want to build a classifier that can predict based on a variety of factors in a student's life if they will be diagnosed with depression.

02

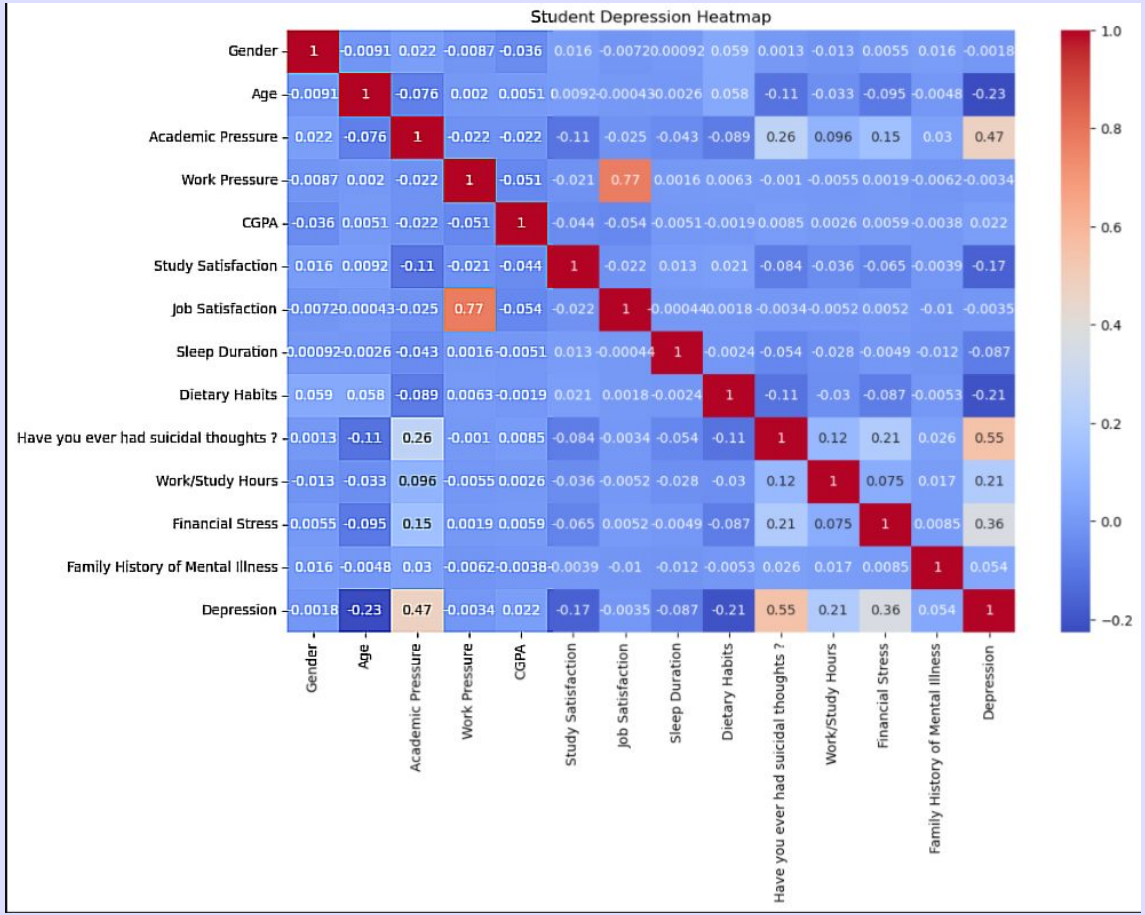
Data Sources

DATA SOURCES

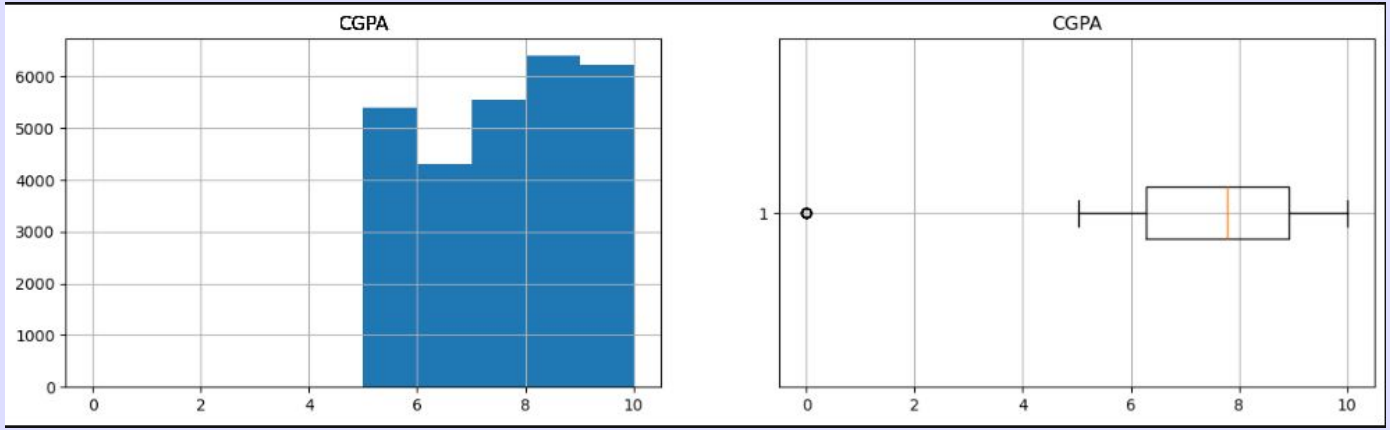
- From Kaggle's Student Depression Dataset
  - 9.41 usability

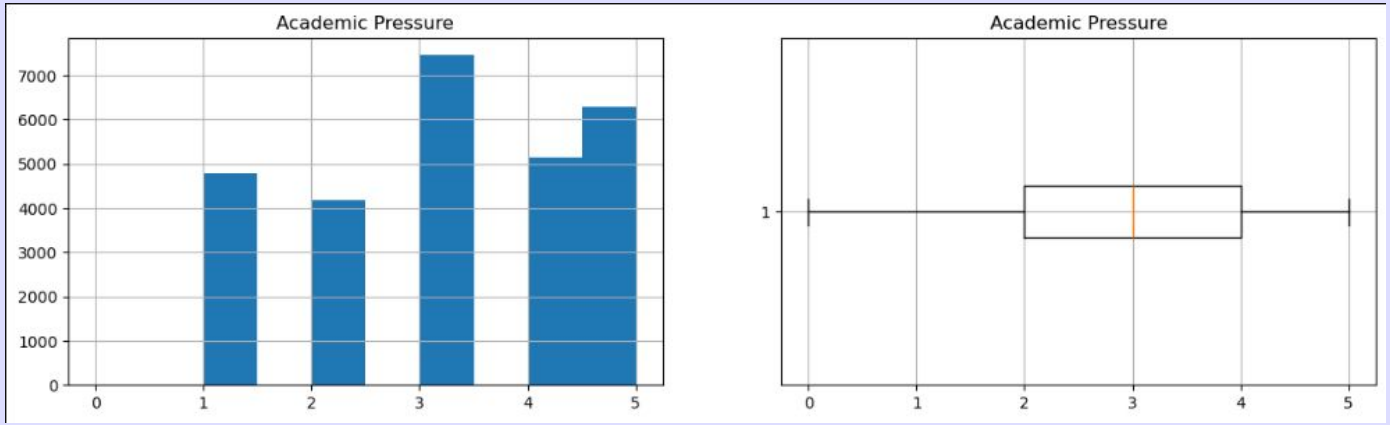
[Student Depression Dataset.]

- Around 28,000 rows, meaning 28k entries of students reporting

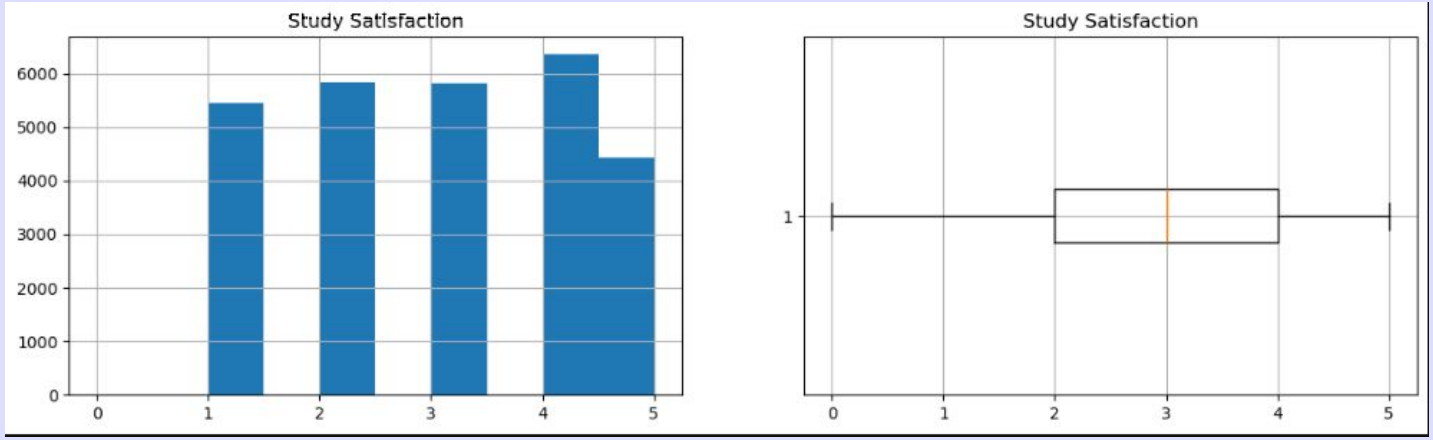


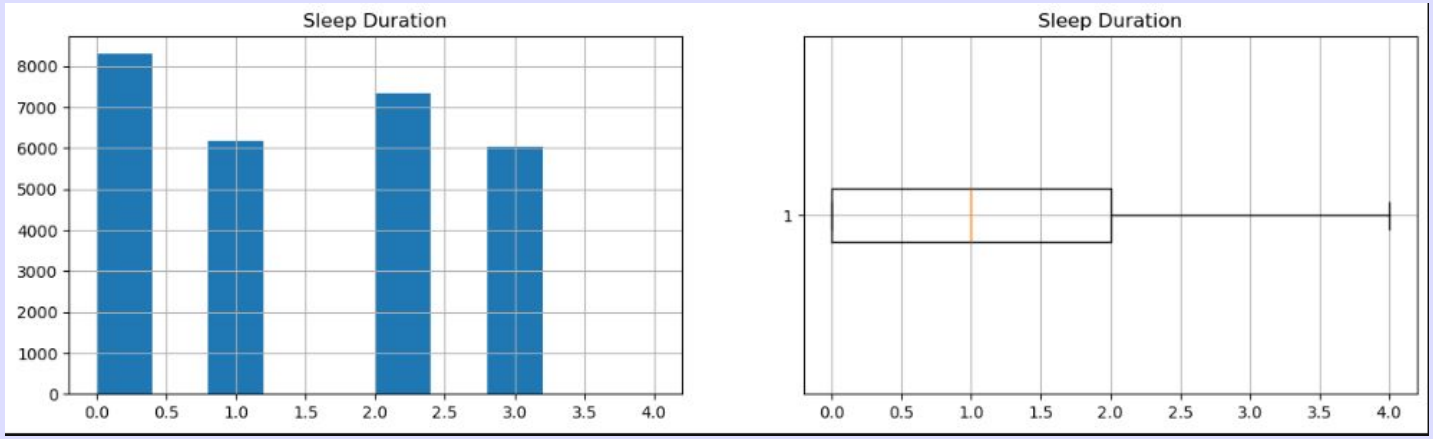
DATA SOURCES



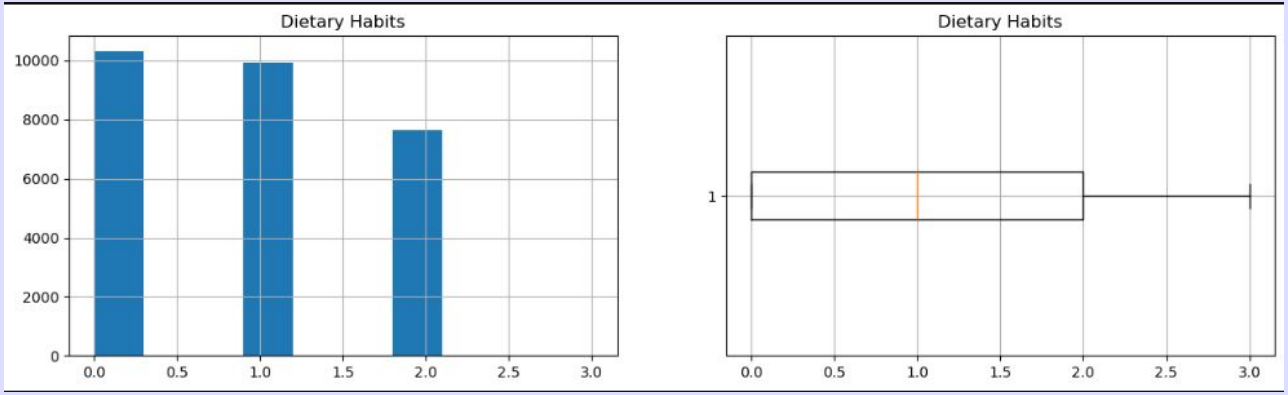


DATA SOURCES





DATA SOURCES



DATASET FEATURES

- Features to consider
  - Gender, Age, Academic Pressure, Work Pressure, CGPA (cumulative GPA), Study Satisfaction, Sleep Duration, Dietary Habits, Have you ever had suicidal thoughts?, Work Study Hours, Financial Stress, Family History of Mental Depression
- Target
  - Diagnosed with depression? Yes or No

PRE-PROCESSING

Discretized for...

- Gender
  - 'Male': 0, 'Female': 1
- Sleep Duration
  - 'Less than 5 hours': 0, '5-6 hours': 1, '7-8 hours': 2, 'More than 8 hours': 3
- Dietary Habits
  - 'Unhealthy': 0, 'Moderate': 1, 'Healthy': 2
- Suicidal Thoughts?
  - 'No': 0, 'Yes': 1
- Family History of Medical Illness?
  - 'No': 0, 'Yes': 1

03

K-Nearest Neighbors Classifier

K-Nearest Neighbors results

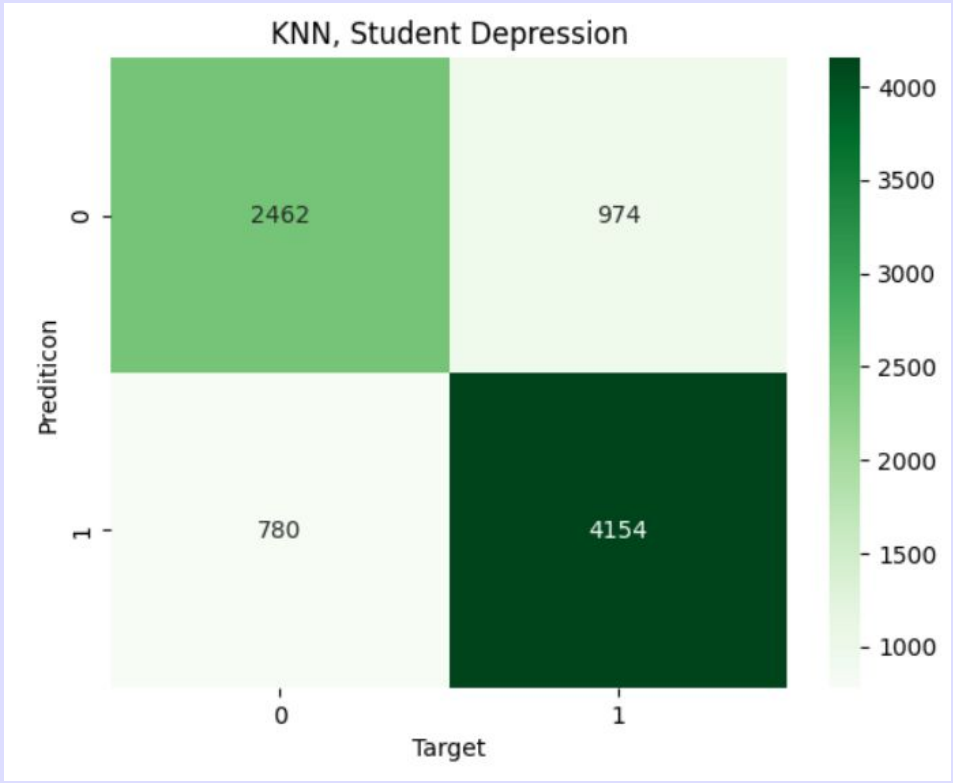
- It classifies a data point based on the majority class of its k-nearest neighbors.
- The choice of k and the distance metric (e.g., Euclidean distance) are important parameters.
- KNN is a classification algorithm used widely in machine learning.

K-Nearest Neighbors results

Results:

- Accuracy: 0.79%

Classification Report:	precision	recall	f1-score	support
0.0	0.76	0.72	0.74	3436
1.0	0.81	0.84	0.83	4934
accuracy			0.79	8370
macro avg	0.78	0.78	0.78	8370
weighted avg	0.79	0.79	0.79	8370

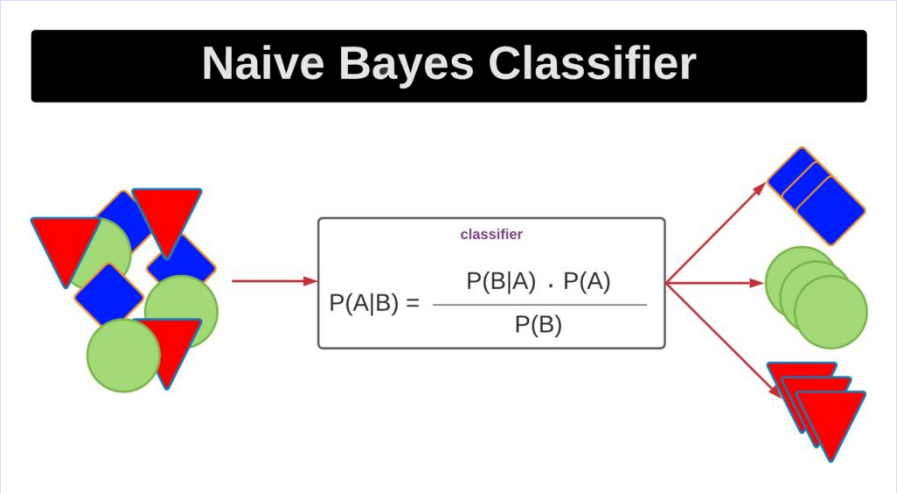


04

Gaussian Naïve Bayes

Gaussian Naïve Bayes

- It relies on Bayes' theorem and assumes that the features describing an observation are conditionally independent when the class label is known.

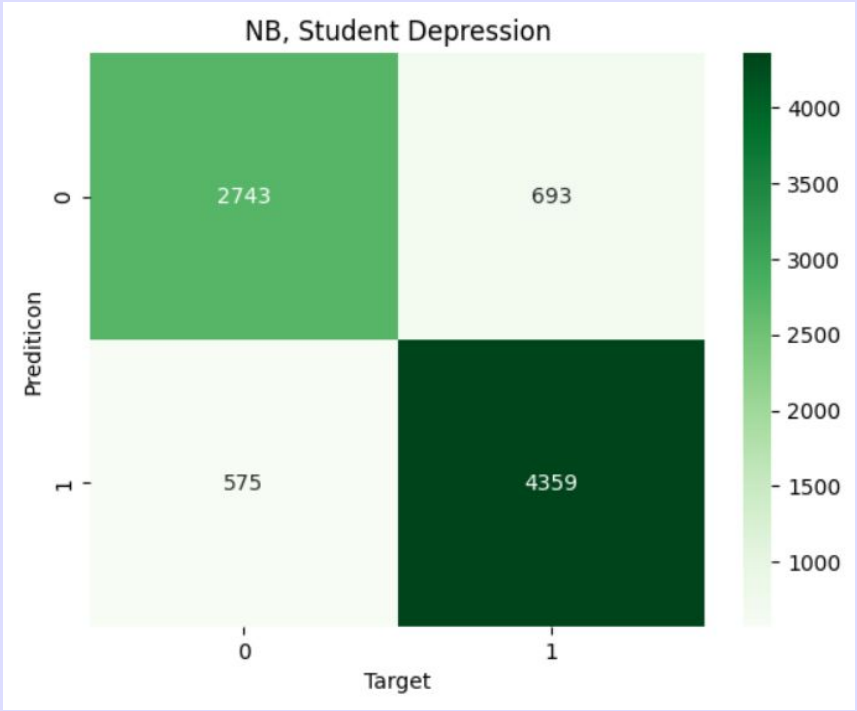


Gaussian Naïve Bayes results

Results:

- Accuracy: 0.85%

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

           0       0.83      0.80      0.81      3436
           1       0.86      0.88      0.87      4934

    accuracy                           0.85      8370
   macro avg       0.84      0.84      0.84      8370
weighted avg       0.85      0.85      0.85      8370
```



05

Support Vector Machine

Support Vector Machine

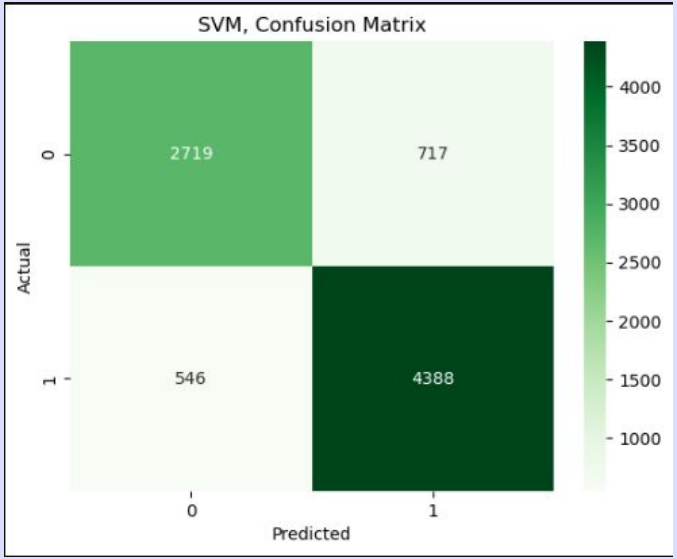
- **Accuracy: 0.85%**
- Identifies patterns in academic, lifestyle, and mental health features
- Helps classify students based on complex interactions among features like "Sleep Duration" and "Dietary Habits."
- Provides a baseline for classification with clear decision boundaries

SVM Model
Accuracy: 0.85

Confusion Matrix:
[[2719 717]
 [546 4388]]

Classification Report:

	precision	recall	f1-score	support
0	0.83	0.79	0.81	3436
1	0.86	0.89	0.87	4934
accuracy			0.85	8370
macro avg	0.85	0.84	0.84	8370
weighted avg	0.85	0.85	0.85	8370



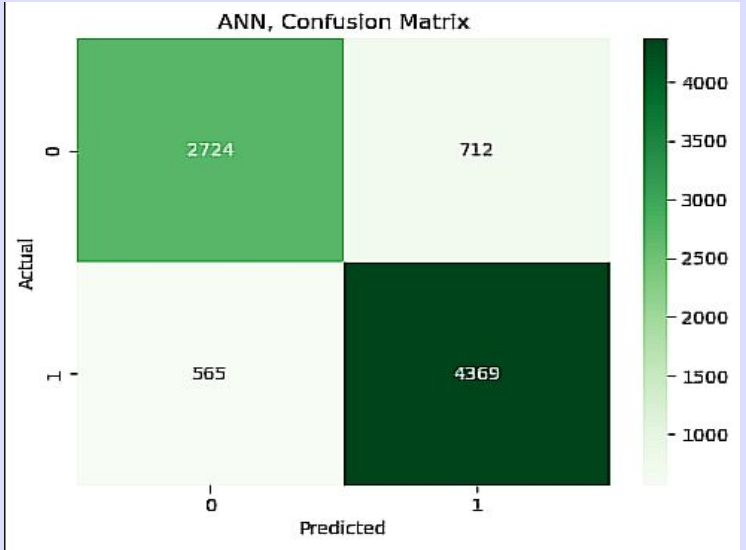
06

Artificial Neural Network

Artificial Neural Network

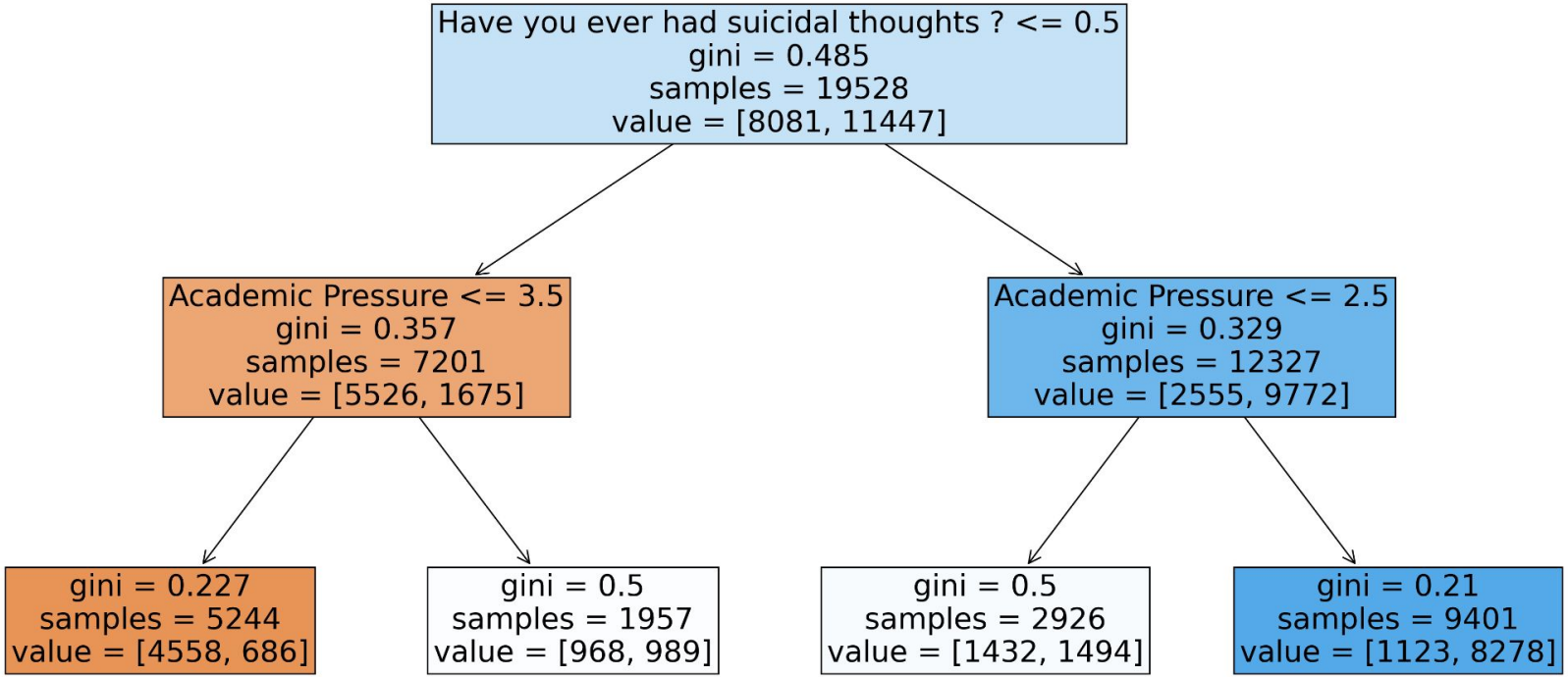
- **Accuracy: 0.85%**
- Learns deeper insights from numerical features like CGPA, Academic Pressure, and Study Satisfaction
- Provides predictions by leveraging interconnections between features influencing depression

```
ANN Model
Accuracy: 0.85

Confusion Matrix:
[[2732  704]
 [ 578 4356]]

Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.80      0.81      3436
           1       0.86      0.88      0.87      4934

    accuracy                           0.85      8370
   macro avg       0.84      0.84      0.84      8370
weighted avg       0.85      0.85      0.85      8370
```



07

Classification and Regression Trees



Tree is pruned, meaning that tree only focuses on significant splits

How CART can explain this dataset

- Easily visualization of splits at important features
- Resistant to outliers
- Feature ranking, makes clear influence certain features have on outcome

```
Accuracy= 0.837037037037037
[[2679  803]
 [ 561 4327]]
              precision    recall  f1-score   support

         0.0       0.83      0.77      0.80      3482
         1.0       0.84      0.89      0.86      4888

    accuracy                           0.84      8370
   macro avg       0.84      0.83      0.83      8370
weighted avg       0.84      0.84      0.84      8370
```

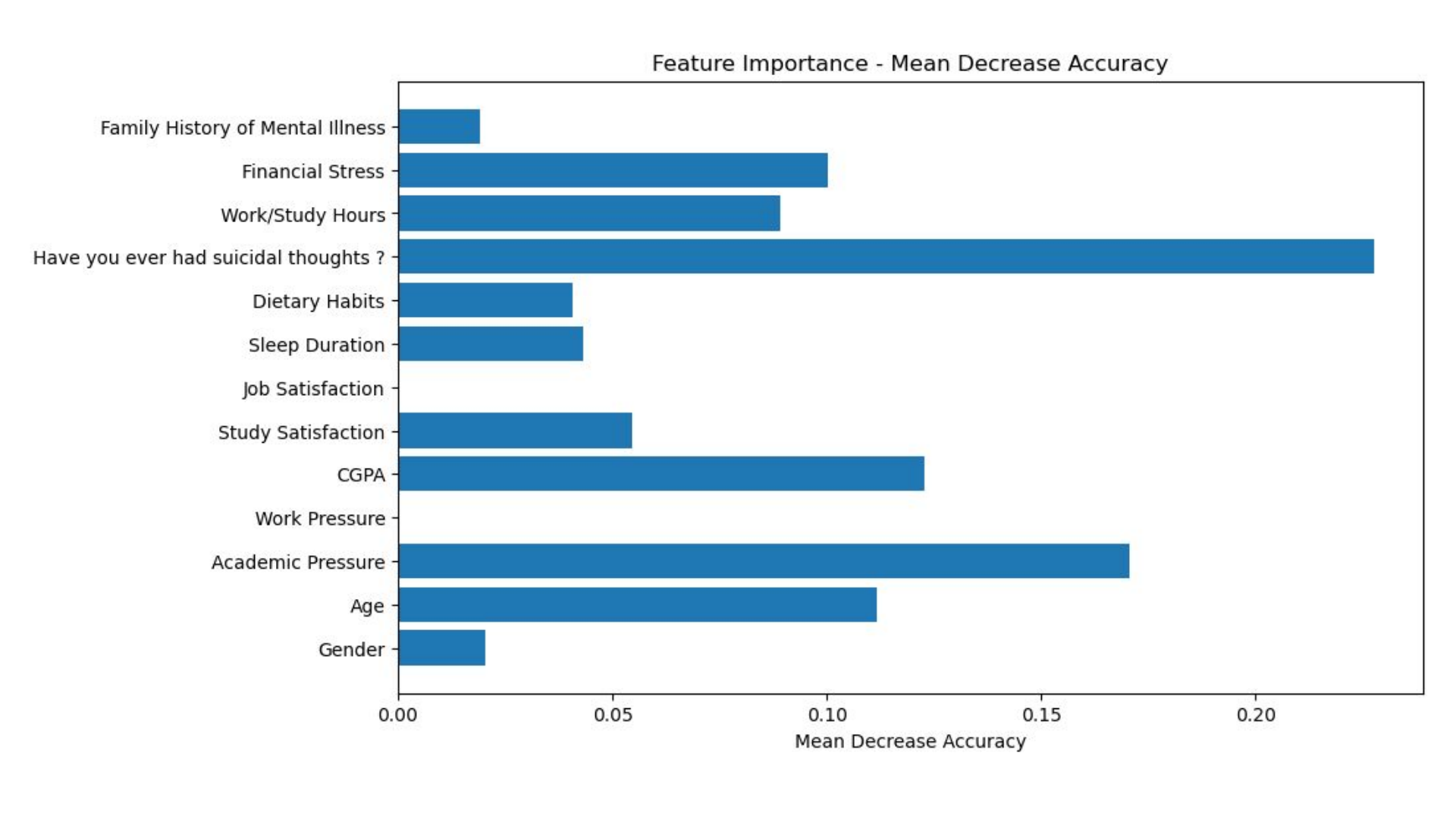
08

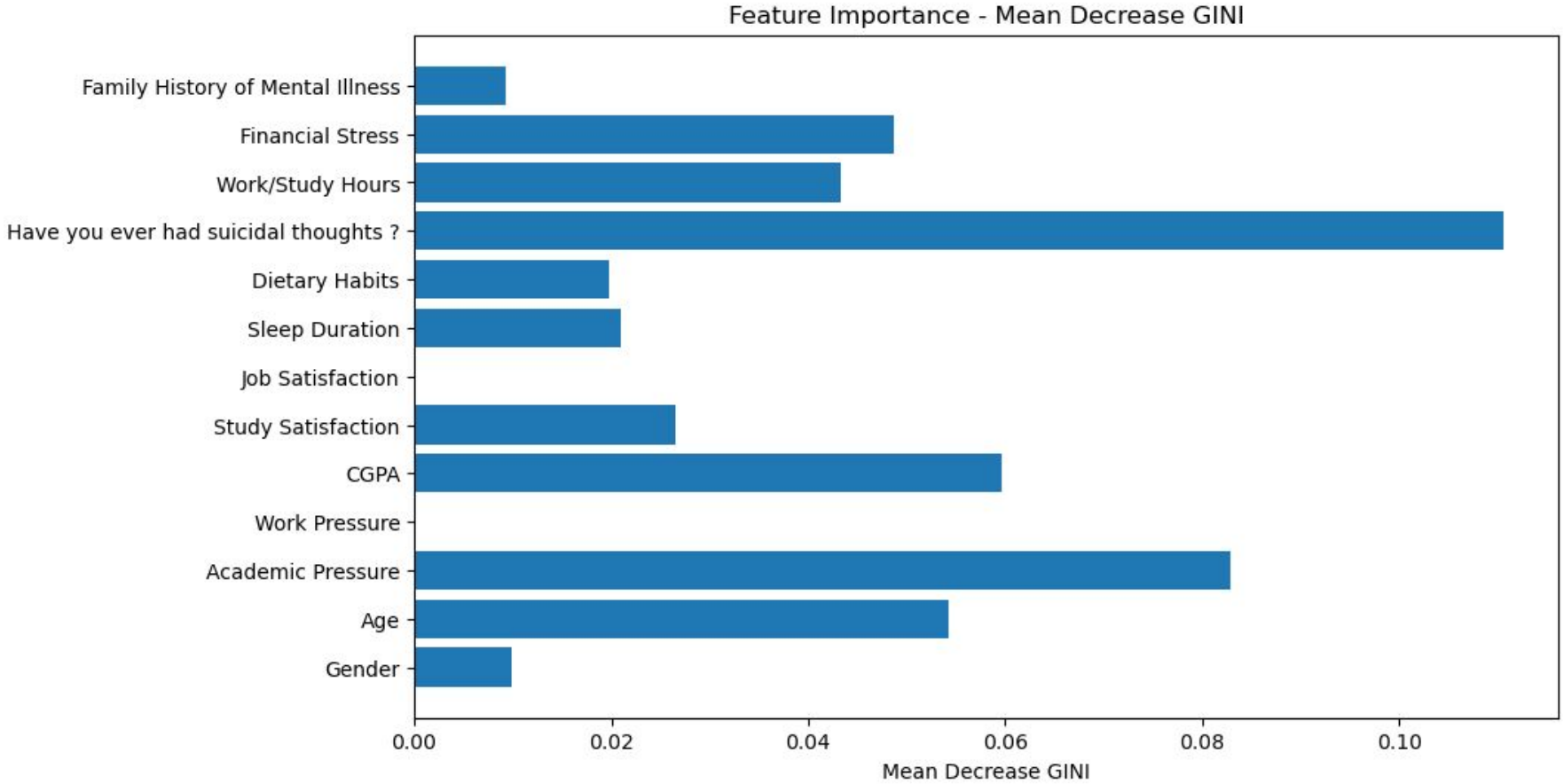
Random Forest

Why is Random Forest classifier appropriate for this dataset?

- Feature ranking, makes clear influence certain features have on outcome
- Reduce overfitting by using averages

```
Accuracy= 0.837037037037037
[[2679  803]
 [ 561 4327]]
              precision    recall  f1-score   support

         0.0       0.83      0.77      0.80      3482
         1.0       0.84      0.89      0.86      4888

    accuracy                           0.84      8370
   macro avg       0.84      0.83      0.83      8370
weighted avg       0.84      0.84      0.84      8370
```


Feature Importance - Mean Decrease Accuracy
Family History of Mental Illness
Financial Stress
Work/Study Hours
Have you ever had suicidal thoughts ?
Dietary Habits
Sleep Duration
Job Satisfaction
Study Satisfaction
CGPA
Work Pressure
Academic Pressure
Age
Gender
0.00
0.05
0.10
0.15
0.20
Mean Decrease Accuracy


Feature Importance - Mean Decrease GINI
Family History of Mental Illness
Financial Stress
Work/Study Hours
Have you ever had suicidal thoughts ?
Dietary Habits
Sleep Duration
Job Satisfaction
Study Satisfaction
CGPA
Work Pressure
Academic Pressure
Age
Gender
0.00
0.02
0.04
0.06
0.08
0.10
Mean Decrease GINI

Feature Scores

Have you ever had suicidal thoughts ?	0.227686
Academic Pressure	0.170638
CGPA	0.122630
Age	0.111651
Financial Stress	0.100072
Work/Study Hours	0.089234
Study Satisfaction	0.054703
Sleep Duration	0.043063
Dietary Habits	0.040740
Gender	0.020372
Family History of Mental Illness	0.019121
Job Satisfaction	0.000084
Work Pressure	0.000005

Results

The performance of all Models:

The most accurate model was: SVM - 0.85%

Why SVM Performs Best:
- Handles high-dimensional features effectively.
- Robust to outliers and noise in data.
- Performs well with scaled numerical features.
- Suitable for relatively balanced datasets.

	Accuracy
SVM	0.849104
NB	0.848507
ANN	0.847431
RF	0.837037
KNN	0.790442
CART	0.777419

References

Dataset: *Student Depression Dataset*

Topic Research: *Depression in Students: Symptoms, Causes, What to Do | Psych Central*

Questions?

THANK YOU