# **Project Proposal**

### Project Title**: Predicting Student Depression Risk**

### Student Name: **Sevastyan Olevinskiy**

## **Project Summary**

This project aims to develop a model to determine **whether a student is likely to suffer from depression symptoms based on academic, behavioral, and lifestyle features**. Using data from a Kaggle dataset, “Student Depression Dataset”. The goal is to determine the most predictive features of the dataset and build a tool that could help in the early detection of students at risk. This would be useful in educational institutions, where such a model could help improve students’ mental health by offering counseling and support to those at risk.

## **Proposed Technical Approach**

## First, I will clean and prepare the dataset for modeling. This will include converting text values, like sleep duration, to numerical values. I will also study the correlation between variables to pick the features that are the most important in predicting depression.

## Then I will train several classification models to predict whether a student is likely to be depressed and decide on the best one. I plan to use decision trees and random forests because they are easy to interpret.

## Finally, I will build a basic user interface, likely using PyEasyGUI, where a user can enter student data and receive a depression risk prediction.

## **Data Sets**

For this project, I will use the **Student Depression Dataset** available on Kaggle: https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset.

The dataset contains information about students, including their age, gender, city, GPA, study habits, sleep duration, dietary habits, financial stress, and other lifestyle factors. The target variable is “Depression,” which shows whether a student is experiencing depression symptoms (1) or not (0).

I will focus on cleaning the data (like converting categories to numbers) and selecting important features for the machine learning models. Since the dataset already has clear labels, I will treat this as a supervised classification problem.

Right now, I am planning to use just this one dataset. If needed later, I may add a second dataset, but it is not required for the project.

## **Experiments and Evaluation**

I’ll shuffle the data and split it roughly 80% for training and 20% for testing. On the training set, I’ll do 5-fold cross-validation to get a feel for how each model performs.

To see how well the models work, I’ll check:

* **Accuracy**: overall percent correct
* **Precision**: how many of the students flagged as depressed really are depressed
* **Recall**: how many of the true depressed cases I actually catch

I’ll also look at the **F1-score**, which balances precision and recall. Then I’ll compare my decision tree and random forest on those numbers and pick the one that performs best on the test set.

At the end, I’ll report the final scores and note which features the winning model found most useful for spotting students at risk.

## **Software**

I will likely use the following software in my project:

* **Python 3**
* **Pandas & NumPy**
  + For basic manipulation with my data.
* **Scikit-learn**
  + For machine learning, including decision trees, random forests and cross-validation.
* **PyEasyGUI**
  + For basic user interface.
* **Matplotlib**
  + To quickly plot the data.
* **GitHub**
  + To manage and share my project easily.

## **Milestones**

* Clear and prepare the data.
* Determine the most influential variables
* Train Decision Tree and Random Forest models
* Validate the models with 5-fold cross-validation.
* Compare the metrics like the accuracy, precision, recall and the F1 score.
* Determine the best model.
* Build a simple user interface, likely using PyEasyGUI.
* Final tests.