Final Project Report:
Student Depression Analysis

ECEG 478: Machine Learning

Khanh Cao

Introduction

This report was made for Bucknell University's ECEG 478 course's Machine Learning final project. This report will outline the student's general plan for the project. The student will utilize the "Student Depression Dataset" accessible through Kaggle and develop a model that can predict a student's risk of depression based on their demographic, academic pressure, GPA, etc. This project aims to raise awareness as well as study what factors most contribute to students' mental health.

The data:

The dataset utilized for this project is the "Student Depression Dataset" by Adil Shamim on Kaggle, which can be publicly accessed through the link:
https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset/data
Per the data's author: This data set is designed for research, and adhered to strict ethical standards with an emphasis on privacy, informed consent, and anonymization.

Data description:

The dataset contains various features grouped into the following categories:

- **Format:** CSV (each row represents an individual student)
- **Features:**
  - **ID:** Unique identifier for each student
  - **Demographics:** Age, Gender, City
  - **Academic Indicators:** CGPA, Academic Pressure, Study Satisfaction
  - **Lifestyle & Wellbeing:** Sleep Duration, Dietary Habits, Work Pressure, Job Satisfaction, Work/Study Hours
  - **Additional Factors:** Profession, Degree, Financial Stress, Family History of Mental Illness, and whether the student has ever had suicidal thoughts
- **Target Variable:**
  - **Depression_Status:** A binary indicator (0/1 or Yes/No) that denotes whether a student is experiencing depression

Dataset Size and Label Domain:

Each data item (i.e., each student) is represented as a **vector of fixed length**, where each element in the vector corresponds to a specific feature or label.

There are **27,901** total records of unique data items. The target label set for depression prediction is the Depression_Status variable, which is binary and finite, with 2 possible values of 0 (Not Depressed) and 1 (Depressed).

The data set is moderately balanced, with 16336 (58.55%) records reporting 1 (Depressed) and 11,565 (41.45%) records reporting 0 (Not Depressed).

Python Modules for Reading the Dataset:

The dataset is provided in CSV format.

There are numerous Python modules capable of reading and handling CSV files, one popular module that can be utilized is the **pandas** library. The student has successfully installed and used pandas before.

Data Interpretation:

Each row in the dataset corresponds to one student and contains both categorical and numerical features. Features like Gender, City, and Profession are categorical and may need label encoding. Features like CGPA, Sleep Duration, and Work/Study Hours are numerical and can be used directly after appropriate normalization or standardization.

The **Depression_Status** variable is the label that models will attempt to predict using the other features as input. Binary classification algorithms will be trained to distinguish between Depressed and Not Depressed statuses.

Data Division into Sets:

- Training Set (70%): The largest portion of the dataset will be allocated to training the model.
- Development Set (15%): For tuning and validating the model.
- Testing Set (15%): For the final evaluation of the model, measures accuracy and performance.

This standard ratio ensures proper model evaluation and avoids overfitting.

Data samples:

Data sample size was condensed for clarity.

id	Gender	Age	City	Study Satisfaction	Job Satisfaction	Sleep Duration	Dietary Habits	Depression
2	Male	33	Visakhapatnam	2	0	'5-6 hours'	Healthy	1
8	Female	24	Bangalore	5	0	'5-6 hours'	Moderate	0
26	Male	31	Srinagar	5	0	'Less than 5 hours'	Healthy	0

Data Preprocessing and Cleaning

Dataset is too large to manually check for missing/invalid values, so exceptions and error handling will be implemented in the code, using try/except blocks for example.

The dataset contains both categorical, numerical, and binary features with different ranges, so renormalization of data will be necessary.

To convert categorical variables into a format suitable for machine learning models:

- Binary Categorical Variables (e.g., Gender, Suicidal Thoughts): These will be encoded as 0 or 1.
- Multi-Class Categorical Variables (e.g., City, Profession): These will be encoded using either one-hot encoding or ordinal encoding, depending on the context and the model used.

The scikit-learn library will be used for encoding to ensure compatibility with downstream models.

To rescale and renormalize data:

All numerical features (e.g., CGPA, Sleep Duration, Work/Study Hours, Job Satisfaction, Financial Stress) will be scaled using either:

- Min-Max Normalization to rescale values to a [0, 1] range, or
- Standardization (z-score normalization) to achieve zero mean and unit variance.

The MinMaxScaler and StandardScaler utilities from scikit-learn can be applied for this purpose.

These steps will ensure that the dataset is clean, consistent, and ready for effective machine learning model training and evaluation.

Representativeness and Potential Bias

The dataset used in this project consists exclusively of university students from various cities across India. It is assumed that all participants are enrolled in Indian universities. The dataset includes students pursuing a wide range of degrees, and the recorded attributes suggest variation in age and socioeconomic background. This diversity indicates a reasonable attempt to reflect the heterogeneity within the Indian university student population.

Limitations:

- **Geographical:** While students are drawn from numerous cities, the dataset is geographically limited to India. As such, findings and models derived from this data may not generalize to students in other countries or cultural contexts, where academic systems, social norms, mental health stigmas, and support structures differ significantly.

- **Balance:** While the dataset appears to include a broad range of values for age, degree programs, and financial stress indicators, suggesting variation in both educational and socioeconomic background. Further analysis will be made to confirm whether this diversity is balanced, or is skewed towards a particular age group, in particular.

- **Self-Report Bias**: Variables such as suicidal thoughts, job satisfaction, and financial stress are self-reported, which introduces the possibility of underreporting or misreporting due to stigma, misunderstanding, or recall errors.

Project Goal:

The primary goal of the project is to develop a binary classification system using machine learning to detect and predict depression in students based on demographic, academic, and lifestyle factors.

This project applies supervised machine learning, with a focus on traditional classifiers like logistic regression and neural networks. If developed into a product, this system could support early mental health intervention by helping educational institutions identify at-risk students.

Computational Setup
The project is implemented in Python 3.10+ using the following libraries:

- Data handling: pandas, numpy
- Modeling: scikit-learn, tensorflow or pytorch
- Preprocessing: sklearn.preprocessing, imbalanced-learn
- Visualization: matplotlib

Methods and Experimental Setup

The workflow includes:

- Data Cleaning and Preprocessing: Handling missing values, encoding categorical variables, and scaling features.
- Data Splitting: Dividing the dataset into training (70%), validation (15%), and test (15%) sets.
- Model Training:
    - Baseline models: Logistic Regression
    - Neural network
- Evaluation Metrics: Accuracy

Summary

This project uses computer-based learning methods to predict whether university students may be experiencing depression. By analyzing information like academic performance, sleep habits, stress levels, and family background, the model helps identify patterns linked to mental health struggles. The goal is to support early detection and offer insights that could improve student wellbeing and guide mental health support efforts in educational settings.