Mental Health Risk Prediction for College Students

PROJECT REPORT

Team Details :

Team Name **UDAAN ML**

Team Members
Vimal Kumar Yadav (10273)
Aryen Mukundam (10198)
Mohit Kumar (10622)
Adhyayan Gupta (10055)
Aryan Jakhar (10305)

---

1. Problem Statement and Initial Approach

Problem Statement

College students face a unique combination of academic, social, and financial pressures, making them particularly vulnerable to mental health challenges like stress, anxiety, and depression. Many students suffer in silence, and institutions often lack the tools to identify those at high risk before their well-being significantly deteriorates.

The objective of this project is to develop a machine learning model that can **proactively predict whether a student is at a "High Risk" or "Low Risk"** for mental health issues based on a range of academic, lifestyle, and demographic factors. By identifying at-risk individuals early, educational institutions can offer timely support and resources, fostering a healthier campus environment.

Initial Approach

Our initial approach was to frame this as a **binary classification problem**. The plan was to:

1. Utilize an open-source dataset containing relevant student information.
2. Engineer a target variable (`Risk`) based on existing mental health indicators (stress, anxiety, depression scores).
3. Preprocess the data by handling missing values, encoding categorical features, and scaling numerical data.

4. Explore and evaluate several standard classification models to determine which one provides the best performance for this specific, sensitive use case, where failing to identify an at-risk student (a false negative) is more critical than incorrectly flagging a healthy student (a false positive).

---

2. Data Collection and Challenges

Data Source

The project utilizes a practical, open-source dataset from GitHub, `health.csv`, which contains self-reported data from college students. The dataset consists of **7023 records** and **21 original features**.

Dataset Description

The features provide a holistic view of a student's life, encompassing academics, personal well-being, and social factors.

Feature	Description
Age	Student's age
Gender	Male/Female
CGPA	Academic performance
Stress_Level	Self-reported stress level
Depression_Score	Self-reported depression score
Anxiety_Score	Self-reported anxiety score
Sleep_Quality	Good / Average / Poor
Physical_Activity	Low / Moderate / High
Diet_Quality	Good / Average / Poor
Social_Support	Low / Moderate / High
Substance_Use	Frequency of usage
Counseling_Service_Use	Frequency of counseling service use
Financial_Stress	Numeric indicator

Residence_Type	On-Campus / Off-Campus / With Family
Risk	Derived binary label (target variable)
...and other features like Course, Relationship Status, etc.	

Challenges Encountered

1. **Missing Data:** The `CGPA` feature contained missing values, which needed to be imputed to avoid data loss. We chose to fill these with the mean of `CGPA` (3.49) according to their respective courses as it was a reasonable assumption for a student's academic standing.
2. **Subjectivity of Data:** The core mental health metrics (`Stress_Level`, `Depression_Score`, `Anxiety_Score`) are self-reported. This introduces a degree of subjectivity and potential bias, as students might under-report or over-report their symptoms.
3. **Categorical Data Handling:** The dataset contained a mix of ordinal and nominal categorical variables, requiring different encoding strategies (label encoding for ordinal, one-hot encoding for nominal) to be used by the machine learning models.

---

3. Methodology and Solution Approach

Data Preprocessing and Feature Engineering

To prepare the data for modeling, the following steps were executed:

1. **Missing Value Imputation:** Missing `CGPA` values were filled with the dataset's mean CGPA of **3.49**. Rows with any other missing values were dropped to ensure data quality.
2. **Dropping Rows with Nulls:** Some rows had null values for `Substance_Use`, which were subsequently dropped to maintain data integrity and model accuracy.
3. **Encoding Categorical Variables:**
    - **Ordinal Features** like `Sleep_Quality` and `Diet_Quality` were mapped to a numerical scale (e.g., Poor=1, Average=2, Good=3).
    - **Binary Features** (`Gender`, `Family_History`) were converted to 0 and 1.
    - **Nominal Features** (`Course`, `Relationship_Status`, `Residence_Type`) were one-hot encoded to create binary columns for each category without implying an order.

3. Feature Engineering (`Risk` Target Variable): A crucial step was creating the binary target variable, `Risk`. A student was classified as "High Risk" (1) if their score for stress, depression, or anxiety was above a threshold of 3. Otherwise, they were "Low Risk" (0).

```
df['Risk'] = df.apply(lambda x: 1 if (x['Stress_Level'] > 3 or
                                      x['Depression_Score'] > 3 or
                                      x['Anxiety_Score'] > 3) else 0, axis=1)
```

This resulted in an imbalanced dataset, with **4265 High-Risk** students (60.8%) and **2742 Low-Risk** students (39.2%). This imbalance informed our choice of evaluation metrics later on.

4. **Feature Scaling:** For distance-based models like KNN and regression models, numerical features (`Age`, `CGPA`, etc.) were standardized using `StandardScaler` to have a mean of 0 and a standard deviation of 1.

Exploratory Data Analysis (EDA) — Key Insights

- A significant majority (~60%) of students in the dataset fall into the "High Risk" category, highlighting the prevalence of mental health concerns.
- Strong correlations were observed between higher stress, anxiety, and depression levels and lifestyle factors like **low sleep quality** and **poor diet**.
- Students with **low physical activity** or experiencing **high financial stress** also demonstrated a greater tendency towards being high-risk.
- While not dominant, features like `Course` (Engineering) and `Residence_Type` (Off-Campus) showed a mild correlation with higher stress levels.

---

4. Model Performance Analysis

Four different classification models were trained and evaluated. The primary goal was to find a model that maximizes **Recall** (correctly identifying high-risk students) without excessively sacrificing **Precision** (avoiding false alarms). The **F-beta score (with beta=2)** was used as a key metric, as it weighs recall twice as much as precision, aligning with our problem's priority.

Model 1: Linear Regression (as a Classifier) (Github Link)

- **Intuition & Approach:** While primarily a regression algorithm, Linear Regression can be adapted for classification by predicting a continuous "risk score" and applying a threshold (0.5) to classify the outcome. It serves as a simple baseline.
- **Performance:**
  - **Accuracy:** 62.9%
  - **Recall (High Risk):** 0.89
  - **Precision (High Risk):** 0.63
- **Model Analysis :** The multimodel classification approach achieved an accuracy of 62.9%, with a recall of 0.89 and precision of 0.63 for the High Risk class. The model

effectively identifies most high-risk cases, minimizing missed detections, though some low-risk instances are misclassified as high risk. Overall, it performs well for high-risk detection, with potential to improve precision through threshold tuning or class weighting.

Model 2: Logistic Regression (Github Link)

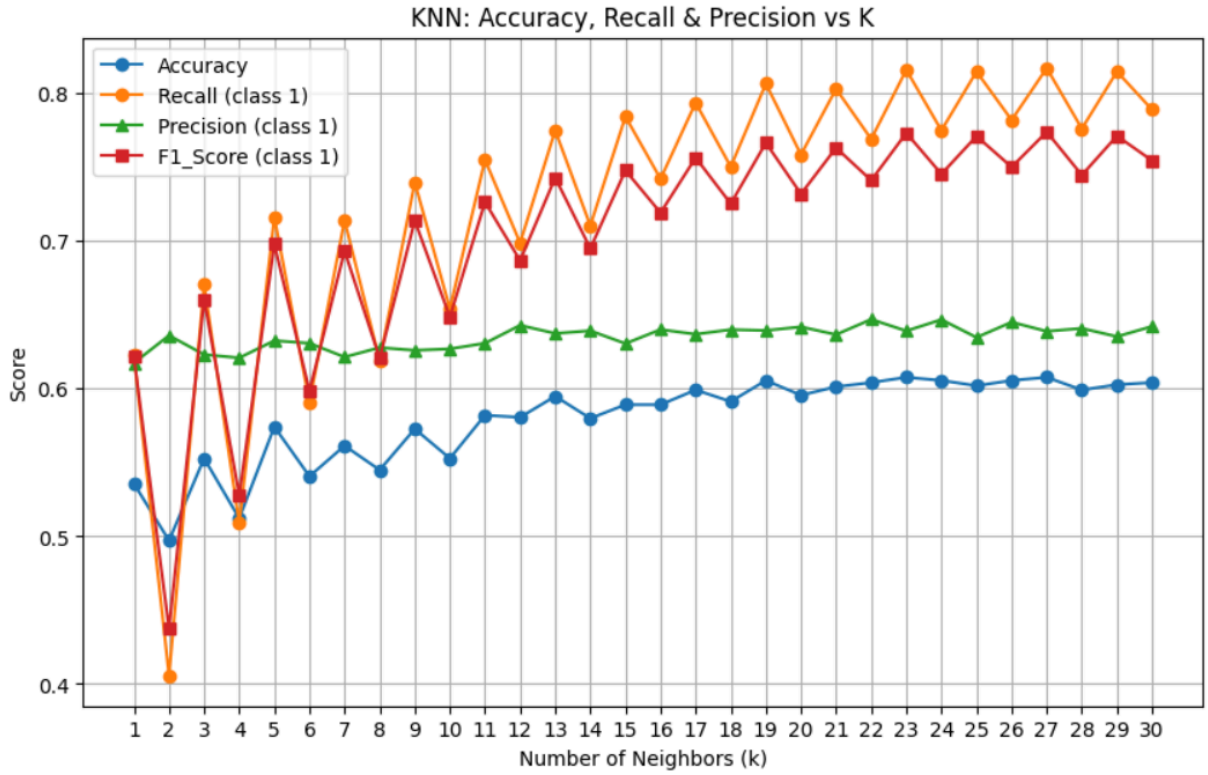
- **Intuition & Approach:** This is a standard, interpretable algorithm for binary classification. We used `class_weight='balanced'` to handle the data imbalance. The key step was **threshold tuning**. By analyzing the Precision-Recall curve, we identified that the default 0.5 threshold was not optimal for maximizing recall.
- **Performance (Optimal Threshold = 0.36):**
  - **Accuracy:** 62%
  - **Recall (High Risk):** 0.95
  - **Precision (High Risk):** 0.62
  - **F2-Score:** 0.86
- **Analysis & Reason for Rejection:** The model achieved an excellent recall of 0.91, meaning it successfully identified 91% of high-risk students. However, its precision of 0.63 was relatively low, leading to a high number of false positives. We sought a model with a better balance.

Model 3: Random Forest (Github Link)

- **Intuition & Approach:** As a powerful ensemble method, Random Forest is robust, handles complex feature interactions, and is not sensitive to feature scaling. We used `GridSearchCV` to find the optimal hyperparameters (`max_depth=12`, `min_samples_leaf=3`).
- **Performance:**
  - **Accuracy:** 62%
  - **Recall (High Risk):** 0.85
  - **Precision (High Risk):** 0.65
  - **F1-Score:** 0.73
- **Analysis & Reason for Rejection:** The Random Forest model provided a good balance between precision and recall, outperforming the baseline linear model and offering better precision than the tuned Logistic Regression. It was a very strong candidate. However, its recall was lower than the tuned KNN model, which was our primary metric.

Model 4: K-Nearest Neighbors (KNN) - Final Chosen Model (Github Link)

- **Intuition & Approach:** KNN is a simple, instance-based model that classifies a data point based on the majority class of its 'k' nearest neighbors. Its effectiveness depends on the feature space where similar students are grouped together.
- **Implementation & Tuning:**
  - **Finding Optimal K:** We iterated through k-values from 1 to 30 and plotted the performance metrics. The plot showed that metrics stabilized and performed well around **k=23**.



    - **Threshold Tuning:** Similar to Logistic Regression, we adjusted the classification threshold from the default 0.5 to **0.4**. This was done to prioritize recall.
- **Performance (k=23, Tuned Threshold = 0.4):**
    - **Accuracy:** 61.3%
    - **Recall (High Risk):** 0.94
    - **Precision (High Risk):** 0.62
    - **F2-Score:** 0.85
- **Analysis & Reason for Selection:** By lowering the threshold, we significantly boosted the recall to **94%**, the highest among all models. While this came with a slight drop in precision to 62%, the trade-off is highly favorable for this problem. The F2-score of **0.85** was also the best, confirming that this model provides the optimal balance for our goal of minimizing false negatives.

Model Performance Summary

Model	Accuracy	Recall (High Risk)	Precision (High Risk)	F2-Score
Linear Regression	62.90%	0.89	0.63	0.79
Logistic Regression	62.10%	0.95	0.62	0.83
Random Forest	62.00%	0.85	0.65	0.79
KNN (Final Model)	**61.30%**	**0.94**	**0.62**	**0.85**

---

5. Assessment of Model Success

The project successfully developed a model capable of identifying students at high risk of mental health issues. The final **KNN model achieves a recall of 94%**, meaning it correctly flags 94 out of every 100 high-risk students.

This high recall rate directly addresses the primary objective: **to minimize the number of at-risk students who go unnoticed**.

The model's precision is **62%**, which implies that for every 100 students flagged as "High Risk," about 38 are false positives. While not perfect, this is an acceptable and manageable trade-off. From an institutional perspective, it is far more desirable to extend support services to a few students who may not critically need them than to miss a single student who does. Therefore, the model is considered a successful solution to the problem statement.

---

6. Next Steps

The current model serves as a strong proof-of-concept. The following steps are for future improvement and implementation:

1. **Data Enhancement:**
    - **Collect More Data:** Expanding the dataset would improve model robustness and generalizability.
    - **Incorporate Longitudinal Data:** Tracking students' mental health and lifestyle metrics over multiple semesters could reveal temporal patterns and lead to more accurate, dynamic predictions.

    Deployment and Practical Application:

    - **Develop a User-Friendly Tool:** The model could be integrated into a simple web application for use by university counseling centers or student affairs departments.
    - **Ethical Implementation:** Create strict protocols for how the model's predictions are used. The goal should be to offer support confidentially and voluntarily, not to label or penalize students. The tool should be a means of outreach, not diagnosis.
2. **Refine the "Risk" Definition:**
    - Collaborate with mental health professionals to refine the criteria for the `Risk` variable, potentially creating a multi-class system (e.g., Low, Moderate, High Risk) for more nuanced interventions.