This project provides a comprehensive machine learning pipeline designed to predict depression risk and integrate those predictions into a proactive mHealth (mobile health) framework.

Project Overview: AI-Driven Depression Risk Prediction

1. Objective

The primary goal was to develop and validate machine learning models capable of identifying individuals at risk for depression based on demographic, lifestyle, and professional factors. The project extends beyond simple classification by providing a scalable framework for real-time intervention within mobile applications.

2. Dataset & Features

The analysis used a dataset of **2,556 participants**, covering a wide range of features:

- **Demographics:** Age, Gender, City, and Education Level.
- **Stressors:** Academic Pressure, Work Pressure, Financial Stress, and Work/Study Hours.
- **Well-being Indicators:** Job/Study Satisfaction, Sleep Duration, and Dietary Habits.
- **Clinical Risk Factors:** Family history of mental illness and history of suicidal ideation.

3. Methodology

- **Feature Engineering:** We unified student and professional data by creating composite scores like Total_Pressure and Total_Satisfaction. We also binned Age into categorical groups for better demographic analysis.
- **Model Selection:** We implemented and compared two distinct algorithms:
  - **Logistic Regression:** Chosen for its high interpretability and ability to provide a "probability score" essential for medical screening.
  - **Random Forest:** Chosen for its ability to capture complex, non-linear interactions between lifestyle factors (e.g., how lack of sleep amplifies the impact of work pressure).
- **Validation:** We used **Stratified 5-Fold Cross-Validation** to ensure the models were stable across different subsets of data, achieving an **F1-score of ~97.6%** with Logistic Regression.

4. Key Findings & Insights

- **Primary Predictors:** The most influential factors driving depression risk in the models were **Financial Stress**, **Suicidal Ideation**, and **Total Work/Study Hours**.
- **Model Performance:** Logistic Regression proved to be a highly effective and stable baseline for this dataset, outperforming the Random Forest in terms of consistency across validation folds.
- **The "Probability" Factor:** Using Logistic Regression allows the system to move away from a binary "Yes/No" and toward a **Risk Continuum**, which is more useful for early intervention.

5. The mHealth Integration Framework

The project concluded by preparing the models for **real-world deployment**:

- **Model Export:** We saved the models using joblib along with a **Metadata Contract** (feature_metadata.json). This ensures that a mobile app and the AI model always speak the same "language" regarding data order and labels.
- **Intervention Logic:** We proposed a tiered intervention strategy based on predicted probability:
    - **Low Risk (<30%):** Wellness tips and positive reinforcement.
    - **Moderate Risk (30–70%):** Guided meditation and clinical screening prompts (e.g., PHQ-9).
    - **High Risk (>70%):** Direct connection to counseling or emergency resources.

Conclusion

This project demonstrates that by combining traditional clinical screening logic (like the PHQ-9 found in NHANES) with modern machine learning, we can create a dynamic, privacy-conscious, and highly accurate tool for mental health support. The exported artifacts are now ready to be integrated into a mobile backend for live testing.