Documentation for Mental Health Prediction Model

1. Dataset Preprocessing Steps

The dataset undergoes several preprocessing steps to ensure data quality and compatibility with the machine learning model: (https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey)

- **Handling Missing Values:**
  - Missing values in categorical and numerical columns are either imputed with appropriate statistical measures or dropped based on their impact.
- **Feature Encoding:**
  - Categorical variables like gender, employment_status, and work_interfere are converted into numerical values using one-hot encoding or label encoding.
- **Feature Selection:**
  - Irrelevant columns (such as comments, state, and Timestamp) are removed.
  - The final dataset retains **47 features** after encoding.
- **Scaling:**
  - If required, numerical columns are normalized to improve model efficiency.

2. Model Selection Rationale

The model is selected based on key performance metrics such as accuracy, precision, recall, and F1-score. The rationale for selection:

- **Model Comparison:**
  - Different models like **Logistic Regression, Random Forest, and XGBoost** were tested.
  - The final selection was based on **predictive performance and generalization ability**.
- **Final Model:**
  - The selected model balances **interpretability** and **performance**.
  - It provides robust predictions while being computationally efficient.

3. How to Run the Inference Script

To run the model and make predictions:

Prerequisites:

Ensure you have the necessary dependencies installed:

pip install -r requirements.txt

Steps:

1. **Run the script** using: python predict_mental_health.py
2. **Input the required details** when prompted.
3. **View the prediction output**, which suggests whether treatment is recommended.

4. UI Usage Instructions

Streamlit Interface

The application provides an interactive UI for users to enter details and receive predictions.

How to Use:

1. Set up a file as api_key.py and add your Gemini API key inside it
2. **Launch the Streamlit App**: streamlit run app.py
3. **Fill in the following details**:
   - Name
   - Age
   - Gender
   - Employment Status
   - Other relevant inputs like family_history, remote_work, and mental_health_consequence.
4. **Click "Predict"** to get the results.
5. The system will process the inputs and display:
   - Whether **treatment is recommended** or not.
   - Additional insights generated via **Gemini AI** chatbot.

5. Additional AI Features

1. Gemini AI Chatbot

- Provides AI-generated mental health guidance based on user queries.
- Uses **Google's Gemini Pro** model for natural language understanding.
- Helps users with stress management, anxiety, and general well-being tips.

2. Image Analysis

- Allows users to upload images for AI-based analysis.
- Can detect facial expressions or patterns related to mental health indicators.

- Redirects users to an external **Streamlit-based image processing tool**.

3. Mental Health Chatbot

- Offers a conversational AI for mental health discussions.
- Provides coping strategies based on user input.
- Available through a dedicated chatbot interface.

Report on Mental Health Data Analysis and Model Findings

1. Introduction

Mental health in workplaces is a crucial concern, and early identification of employees requiring support can lead to better well-being and productivity. This project analyzes mental health survey data and builds a predictive model to assess whether an individual might need mental health treatment based on various workplace and demographic factors.

2. Data Analysis and Findings

2.1 Dataset Overview

The dataset contains information on multiple aspects affecting mental health, including:

- **Demographic Factors**: Age, Gender, Country, Employment Status
- **Workplace Factors**: Remote Work, Number of Employees, Mental Health Support Availability
- **Personal Mental Health History**: Family History, Previous Treatment, Work Interference

2.1.1 Data Cleaning and Preprocessing

- **Missing Values:**
    - Some categorical variables contained missing values, which were either imputed or marked as "Unknown."
    - The comments column had excessive missing values and was dropped.
- **Encoding & Feature Engineering:**
    - Categorical variables were transformed using **one-hot encoding**.
    - Features were standardized where necessary for optimal model performance.
    - After encoding, the dataset contained **47 final features**.

2.2 Exploratory Data Analysis (EDA)

2.2.1 Mental Health Treatment Distribution

- Approximately **48% of respondents reported seeking treatment for mental health conditions**, while 52% did not.
- The dataset is relatively balanced, reducing the risk of model bias.

2.2.2 Age vs. Mental Health Treatment

- Employees **aged 25-35 were the most likely to seek treatment**, possibly due to increased work stress and awareness.
- Older individuals (above 50) had a **lower treatment rate**, which might indicate **stigma or lack of awareness**.

2.2.3 Gender Influence on Mental Health

- **Females reported higher treatment rates** compared to males.
- Non-binary and other gender identities also had **higher treatment rates**, possibly due to unique social pressures.

2.2.4 Workplace Factors and Mental Health

- **Remote Work:**
    - Employees working remotely had **slightly better mental health outcomes**, possibly due to flexible work conditions.
- **Company Size:**
    - Employees in large companies (**1000+ employees**) were **less likely to seek treatment**, indicating a possible lack of mental health support in corporate environments.
    - Medium-sized companies (**100-500 employees**) had **higher rates of mental health treatment access**.

3. Model Performance and Findings

3.1 Model Selection

Several models were evaluated for predicting whether an individual needs mental health treatment.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	78.5%	76%	75%	75.5%
Random Forest	82.1%	79%	80%	79.5%
XGBoost	**85.3%**	**83%**	**84%**	**83.5%**

- **XGBoost was chosen** as the final model due to its superior performance.
- The model was trained using **cross-validation** and **hyperparameter tuning** to prevent overfitting.

4. Key Insights from Predictions

4.1 Most Influential Factors in Predictions

Using **feature importance analysis**, the most impactful predictors of mental health treatment were:

- **Work Interference**: Employees who frequently experience work interference due to mental health issues were more likely to need treatment.
- **Company Mental Health Policies**: Lack of proper mental health resources at the workplace significantly increased treatment needs.
- **Family History**: Individuals with a family history of mental illness were at **30% higher risk** of needing treatment.

4.2 Limitations of the Model

- The dataset is **self-reported**, meaning there could be bias in responses.
- Some **features like "Country" were removed** due to inconsistent regional representation.
- The model does not diagnose mental illness but **only predicts treatment necessity based on workplace and personal factors**.

5. Conclusion:

- The analysis highlights the **significance of workplace policies and awareness in improving mental health outcomes**.
- **Employers should focus on providing better mental health support** to employees, especially in larger organizations.