Schizophrenia Prediction Model Architecture

1. Objective

To develop a binary classification model that predicts whether a patient is schizophrenic (1) or not schizophrenic (0) using a dataset of 10,000 patient records with no missing values.

2. Data Input

- Dataset consists of 10,000 rows and 18 features:
    - **Patient Information:**
        - Age (18-80)
        - Gender (0: Female, 1: Male)
        - Education Level (1: Primary, 2: Middle School, 3: High School, 4: University, 5: Postgraduate)
        - Marital Status (0: Single, 1: Married, 2: Divorced, 3: Widowed)
        - Occupation (0: Unemployed, 1: Employed, 2: Retired, 3: Student)
        - Income Level (0: Low, 1: Medium, 2: High)
        - Living Area (0: Rural, 1: Urban)
    - **Medical Information:**
        - Disease Duration (1-40 years for schizophrenia patients)
        - Hospitalizations (0-10 for schizophrenia patients)
        - Family History (0: No, 1: Yes)
        - Substance Use (0: No, 1: Yes)
        - Suicide Attempt (0: No, 1: Yes)
        - Positive Symptom Score (0-100)
        - Negative Symptom Score (0-100)
        - GAF Score (0-100)
    - **Social and Behavioral Factors:**
        - Social Support (0: Low, 1: Medium, 2: High)
        - Stress Factors (0: Low, 1: Medium, 2: High)
        - Medication Adherence (0: Poor, 1: Moderate, 2: Good)

3. Data Preprocessing

1. **Feature Scaling:**
    - Apply Min-Max Scaling for numerical columns like Age, Symptom Scores, and GAF Score to normalize the data between 0 and 1.
2. **Train-Test Split:**
    - Split the dataset into:
        - 80% Training Data
        - 20% Testing Data

4. Model Selection

- **Logistic Regression:**
    - Chosen for its simplicity, interpretability, and suitability for binary classification problems.
    - Assumes a linear relationship between input features and the log-odds of the target variable.

5. Model Training and Prediction

1. **Training Phase:**
    - Fit the logistic regression model using the training dataset.
2. **Prediction Phase:**
    - Predict the probability of a patient being schizophrenic (1) or not (0).

6. Model Evaluation

- **Primary Metric: Accuracy**
    - Evaluate the model's correctness by calculating the proportion of correct predictions:
- **Confusion Matrix:**
    - Display True Positives, True Negatives, False Positives, and False Negatives to analyze performance in detail.

7. Pipeline Summary

1. Input patient data (18 features).
2. Preprocess the data (encoding, scaling, train-test split).

3. Train a logistic regression model.
4. Make predictions on new patient data.
5. Evaluate performance using accuracy and a confusion matrix.