**PART 2: Case Study Application**

**1. Problem Scope**

**Problem Definition**:  
Hospital readmissions within 30 days of discharge are costly, often preventable, and may reflect poor quality of care. The goal is to develop an AI system to predict which patients are at high risk of being readmitted.

**Objective**: To use patient data to generate real-time readmission risk scores at the time of discharge, enabling targeted interventions.

**Stakeholders**:

* Healthcare Providers: Use predictions to optimize discharge planning and follow-ups.
* Hospital Administrators: Reduce readmission rates and associated penalties (e.g., CMS).
* Patients: Benefit from improved care continuity.
* Data Scientists/IT Teams: Develop, deploy, and maintain the model.
* Compliance Officers: Ensure adherence to healthcare laws (e.g., HIPAA).

**2. Data Strategy**

**a. Data Sources**:

* Electronic Health Records : Diagnoses, lab results, procedures, medications, discharge summaries.
* Patient Demographics: Age, gender, ethnicity, insurance type.
* Utilization History: Prior admissions, ED visits, length of stay.

**b. Ethical Concerns**:

1. Patient Privacy: Risk of re-identification from sensitive health data. All data must be de-identified or securely encrypted.
2. Bias and Fairness: Historical data may reflect systemic healthcare disparities, leading to biased predictions against vulnerable populations (e.g., minorities or low-income groups).

**c. Preprocessing Pipeline**:

1. Data Cleaning: Remove duplicates, handle missing values (impute or flag).
2. Feature Engineering:
   * Create flags for chronic conditions (e.g., diabetes, CHF).
   * Generate discharge-to-readmission time interval (if historical readmission data is used).
   * Encode discharge disposition (e.g., home, skilled nursing)-
3. Encoding & Scaling:
   * One-hot encode categorical variables.
   * Normalize numerical features (e.g., lab values).
4. Train-test Split: Stratified sampling to balance readmitted vs non-readmitted cases.

**3. Model Development**

**a. Model Selection**:

* Gradient Boosting Machine (e.g., XGBoost): Performs well on structured healthcare data, handles missing values, and provides interpretability via SHAP values.
* Justification: XGBoost is good to overfitting, effective with imbalanced datasets, and scalable for real-time scoring.