1. Problem Definition

Hypothetical Problem:
Predicting the likelihood of hospital readmission within 30 days after patient discharge.

Unexpected readmissions can indicate inadequate post-discharge treatment or poor patient outcomes, and they are expensive. In order to achieve policy compliance (such as insurance penalties for high readmission rates) and improve the quality of care, hospitals strive to lower readmission rates.

Objectives:

1. For the purpose of follow-up planning, identify high-risk patients at the time of discharge.
2. Enhance post-discharge care with focused interventions (e.g., medication tracking, nurse visits).
3. Cut down on preventable readmissions and related medical expenses.

Stakeholders:

- Hospital administrators are interested in lowering fines and enhancing performance indicators.
- Primary Care Physicians: Determine whether patients require closer monitoring upon discharge by using the model's predictions.

Key Performance Indicator(KPI)
A high-performing model that strikes a balance between avoiding false alarms and identifying the majority of real readmission risks is the Key Performance Indicator (KPI): Readmission prediction accuracy > 90%.

2. Data Collection & Preprocessing

Data Sources:

1. **Electronic Health Records (EHRs)** – Includes diagnoses, procedures, vitals, medications, discharge summaries, and lab results.

2. **Patient Demographic & Socioeconomic Data** – Includes age, gender, income level, insurance type, and caregiver availability.

Potential Bias:

- **Underrepresentation of marginalized groups** (e.g., low-income or rural patients) may lead to poor prediction accuracy for those populations due to lack of sufficient

data or skewed patterns.

Preprocessing Steps:

1. **Imputation of Missing Lab Values** – Fill missing labs using median values or flag as missing.
2. **Feature Engineering** – Create new features like “number of prior hospitalizations” or “comorbidity count.”
3. **Encoding Categorical Variables** – Convert diagnosis codes, insurance types, and discharge locations using one-hot encoding.

3. Model Development

Model Chosen:
Gradient Boosting (XGBoost)

Justification:

- Handles complex patterns in clinical data
- Often achieves high accuracy on structured medical data
- Offers built-in regularization to prevent overfitting
- Feature importance plots can be used for interpretability

Data Splitting Strategy:

- **Training Set (70%)**: Teaches the model the patterns.
- **Validation Set (15%)**: Used for hyperparameter tuning and early stopping.
- **Test Set (15%)**: Measures final performance on unseen data.

Hyperparameters to Tune:

1. **`learning_rate`** – Controls how much each tree corrects previous errors. Lower values improve accuracy but require more trees.
2. **`max_depth`** – Controls the complexity of each decision tree, balancing bias and variance.