**Project Title:** Predicting Hospital Readmissions Based on Patient and Treatment Data

**Topic Chosen:** Hospitals

**Business Understanding**

**Problem Description:** In healthcare systems, readmissions are a major concern because they lead to increased healthcare costs and negatively impact patient outcomes. Hospitals are incentivized to reduce readmissions, especially for specific conditions like heart failure, pneumonia, or surgical procedures. Hospitals seek to understand what factors contribute to patients being readmitted within 30 days of discharge.

Understanding and predicting readmissions can help improve patient care by enabling preventive interventions for high-risk patients. The goal is to identify patients at high risk of being readmitted so that appropriate care plans can be put in place to reduce this likelihood.

**Question:** *Can we predict whether a patient will be readmitted to the hospital within 30 days of discharge based on their medical history, treatment plan, and demographic factors?*

**Analytic Approach**

To answer the business question, a predictive modeling approach will be used. Specifically, classification models will be employed, as the goal is to predict whether or not a patient will be readmitted within 30 days. Some machine learning algorithms may be explored to develop this predictive model.

**Data Requirements**

To predict hospital readmissions, the following data elements will be required:

* **Patient demographic data:** Age, gender, socioeconomic status
* **Medical history:** Pre-existing conditions, number of previous hospitalizations, past surgeries
* **Treatment data:** Type of treatment or procedure received, medications prescribed, length of stay, severity of illness
* **Discharge information:** Discharge instructions, follow-up appointments
* **Post-discharge data:** Information on whether patients complied with post-discharge recommendations
* **Hospital characteristics:** Size, type, and capacity of the hospital, care quality indicators
* **Time-related factors:** Date of discharge and readmission, time between discharge and readmission

**Data Collection**

The required data can be collected from hospital electronic health records (EHR) and claims data. These records typically contain detailed information about the required data. If necessary, external sources such as insurance claims databases or public health data may be used to supplement the dataset with socio-demographic information or community health indicators.

**Data Understanding and Preparation**

Once the data is collected, it will need to be cleaned and preprocessed. This includes:

* **Data Cleaning:** Handling missing values, ensuring consistency in data formats, and correcting errors in the data (e.g., erroneous dates, duplicates).
* **Data Transformation:** Encoding categorical variables (such as gender or type of diagnosis), normalizing or scaling numeric variables like age and length of stay.
* **Exploratory Data Analysis (EDA):** Understanding the relationships between variables, identifying important predictors (such as whether certain treatments or demographics correlate with higher readmission rates).
* **Feature Engineering:** Creating new features, such as the time since last readmission or interactions between treatment type and patient age.

**Modeling and Evaluation**

Multiple machine learning models will be tested to predict whether a patient will be readmitted:

* **Model Selection:** Classification models such as Logistic Regression and Decision Trees will be explored.
* **Model Training:** The data will be split into training and testing sets. The model will be trained on the training data, with the performance evaluated on the testing set.
* **Evaluation Metrics:** Accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC) will be used to assess the model's performance. Since reducing false negatives is important, recall may be prioritized.
* **Model Tuning:** Hyperparameter tuning will be performed to improve model performance, using techniques such as cross-validation and grid search.