1. **Business Understanding → Define the Problem**

* The goal was to **predict hospital readmission** for **congestive heart failure (CHF)** within **30 days**.
* Medical experts helped define **what CHF is** and which diagnosis-related group (DRG) codes represent it.
* The **business objective** was to **reduce hospital readmissions**, **improving patient care and lowering costs**.

**Key Question:** *How can we identify high-risk patients before they get readmitted?*

1. **Analytical Approach → Choose the Type of Analysis**

* Since the goal is to predict a **Yes/No** outcome (readmission or not), this is a **classification problem**.
* A **decision tree classification model** was chosen to predict patient readmission.

**Key Question:** *What type of machine learning approach best fits this problem?*

1. **Data Requirements → Identify Needed Data**

* The team determined they needed **patient hospitalization data** (admission/discharge dates, diagnoses, procedures).
* Other relevant factors like **co-morbidities (diabetes, hypertension), prescriptions, and hospital visits** were also required.

**Key Question:** *What data is necessary to make accurate predictions?*

1. **Data Collection → Gather Raw Data**

* Data came from **multiple sources**:
  + **Hospital claims** (admissions, discharge records, treatments).
  + **Doctor visits** (diagnoses, prescriptions).
  + **Patient demographics** (age, gender, insurance type).
* The dataset contained **multiple records per patient** (transactional format).

**Key Question:** *Where can we find the data needed for the analysis?*

1. **Data Understanding → Assess Data Quality**

* The team analyzed **missing values, invalid data, and inconsistencies**:
  + Checked for **duplicate records**.
  + Identified **outliers** (unrealistic values like negative ages).
  + Standardized **medical codes** for CHF.
* A **literature review** was done to ensure **important medical factors weren’t missing**.

**Key Question:** *Is the data reliable, complete, and relevant for modeling?*

1. **Data Preparation → Clean & Transform the Data**

* The **raw transactional data was aggregated** to create **one record per patient**.
* **Feature engineering** was done to create new variables, such as:
  + **Number of past hospital visits**.
  + **Time since last doctor visit**.
  + **Co-morbidities like diabetes and hypertension**.
* **Missing values were handled** (either removed or filled).
* **Categorical data was converted** for machine learning (e.g., insurance types converted to numbers).
* The final dataset was **structured and ready for modeling**.

**Key Question:** *How do we clean and format the data to improve accuracy?*

1. **Modeling → Build the Prediction Model**

* A **decision tree classification model** was trained using the cleaned dataset.
* The dataset was split into **training and testing sets** to evaluate performance.

**Key Question:** *Which algorithm best predicts patient readmission?*

1. **Evaluation → Validate the Model**

* The model’s accuracy was tested using the **test dataset**.
* If performance was low, **features were refined** to improve results.

**Key Question:** *Does the model perform well, or do we need more improvements?*

1. **Deployment → Use the Model in Practice**

* If successful, the model could be used by **hospitals and doctors** to identify high-risk patients.
* Doctors could **intervene early** to prevent readmissions.

**Key Question:** *How can we use this model in real-world hospital settings?*

**Final Takeaway**

This case study **perfectly follows the Data Science Methodology** from **Business Understanding to Deployment**.

✅ **Data Science is not just coding—it’s a process!**  
✅ **Each step ensures we build an accurate, useful model.**