**1. What’s the goal?**  
We’re working with the **Diabetes 130-US Hospitals dataset**.  
Our goal is to **predict whether a patient will be readmitted within 30 days** after discharge.

**2. How did we prepare the data?**

* First, we set up the **target column**: if a patient was readmitted in less than 30 days, we mark it as 1, otherwise 0.
* We removed ID columns like patient number and encounter ID since they don’t help predictions.
* Then we handled missing values:
  + For numbers, we filled them with the **median**.
  + For categories, we filled them with the **most frequent value**.
* After that, we **scaled the numeric features** so they’re on the same range, and converted categories into numbers using **one-hot encoding** (basically turning them into 0s and 1s).

This gives us a clean dataset ready for machine learning.

**3. How did we train the models?**  
We split the dataset:

* **80% for training** the models.
* **20% for testing**, to fairly evaluate performance.

We tested **three models**:

* Logistic Regression → simple baseline model.
* Random Forest → more powerful, uses many decision trees.
* Gradient Boosting → advanced model that builds trees step by step to improve accuracy.

All these steps — cleaning, scaling, encoding, and model training — were combined into a **pipeline**, so the process is organized and avoids mistakes like data leakage.

**4. How did we evaluate them?**  
We didn’t just look at accuracy — because in medical data, accuracy alone can be misleading.  
We also checked:

* **Precision** → when the model predicts readmission, how often is it right?
* **Recall** → out of all patients who were readmitted, how many did we catch?
* **F1 Score** → balances precision and recall.
* **ROC-AUC** → measures how well the model separates readmitted vs non-readmitted patients.

We also visualized results with **confusion matrices** and **ROC curves** to see errors and trade-offs clearly.

**5. Why does this matter?**

* If the model works well, hospitals can identify high-risk patients early and give them better follow-up care.
* That could mean fewer re-hospitalizations, lower costs, and better patient outcomes.

s

🎤 **Closing line (to wrap up in the meeting):**

“So far, we’ve built a clean pipeline that prepares the data, trains three strong models, and evaluates them with detailed metrics and visualizations. The results are promising, and our next step is to fine-tune, explore stronger algorithms, and move closer to a deployable solution.”

**1️⃣ Logistic Regression (30s):**  
Logistic Regression is a simple model that looks at all patient features, combines them with weights, and calculates the probability of readmission. If the probability is above a threshold, it predicts readmission. It’s fast, easy to understand, and helps show which features increase or decrease the risk. It’s a great baseline model for comparison.

**2️⃣ Random Forest (30s):**  
Random Forest builds many decision trees on random subsets of the data. Each tree makes a prediction, and the final result is based on the majority vote. It can capture complex patterns and interactions between features, like how age and medications together affect readmission. It’s more powerful than logistic regression and is robust to noise and outliers.

**3️⃣ Gradient Boosting (30s):**  
Gradient Boosting builds decision trees one by one, where each new tree focuses on correcting the mistakes of the previous trees. This sequential learning makes it very strong for complex patterns in the data. It usually gives the highest accuracy, especially when predicting complicated outcomes like hospital readmissions