Understanding the Problem

We set out to build a model that can predict which patients are at risk of being readmitted to the hospital. Getting this right is very important — it helps improve patient care, reduce hospital workload, and save costs. We tested several machine learning models and chose the one that gave the most balanced and trustworthy results.

Data Preparation

Here's what we did to get the data ready for modeling:

- Cleaned the data: Removed any missing values and duplicate records.
- Engineered features: Created new variables from existing patient data to make the model smarter.
- Split the data: Divided it into training and test sets so we could evaluate models fairly.
- Scaled and encoded:
  - Scaled continuous features for better performance.
  - Encoded categorical variables to make them machine-readable.

Model Evaluation and Comparison

We used four important metrics to evaluate each model:

- Accuracy: How many predictions were correct overall.
- F1 Score: A balance between precision and recall — crucial because mistakes can be costly in healthcare.
- ROC AUC: Measures how well the model distinguishes between patients who will and won't be readmitted.
- Log Loss: Measures how confident the model is in its predictions (lower is better).

Model Performance:

- **Random Forest:**
  - Accuracy: 88.66%
  - F1 Score: Best among all models (but still low)
  - ROC AUC: 0.6250 (highest)
  - Log Loss: 0.3719 (lowest)
- **Support Vector Machine (SVM):**
  - Accuracy: 88.77% (a little higher than Random Forest)
  - F1 Score: 0 (failed to predict positive cases)
  - ROC AUC: 0.5504
  - Log Loss: 0.4535
- **Decision Tree:**
  - Accuracy: Lower than Random Forest and SVM
  - F1 Score: 0.1706 (better than SVM, but still not great)
  - ROC AUC: 0.5293
  - Log Loss: 7.4700 (very high, unreliable)

Best Model: Random Forest

Based on all the metrics, Random Forest stood out because:

- It gave the most balanced and reliable performance.
- It was best at separating high-risk and low-risk patients.
- It had the highest prediction confidence (lowest log loss).

Why Random Forest Works Well in Healthcare:

- Very Robust: Handles messy and complex healthcare data without overfitting.
- Flexible: Works with both numbers and categories and can manage missing data.
- Insightful: Tells us which factors are most important for predicting readmissions — valuable for doctors and hospitals.

Conclusion and Next Steps

In short, Random Forest is the best choice for predicting hospital readmissions in this project.

✅ Recommendations:

1. Deploy Random Forest as the primary predictive model.
2. Use feature importance to find and monitor key patient risk factors.
3. Integrate the model into hospital systems to proactively manage high-risk patients.
4. Keep updating the model regularly with new data to maintain its accuracy.