**Predicting 30-Day Hospital Readmissions: My Approach and Findings**

This project focused on predicting whether a patient would be readmitted to the hospital within 30 days using a combination of structured data and free-text discharge notes. The dataset included 200 patient records with features like age, gender, diagnosis code, medication type, number of previous admissions, length of stay, and a discharge summary. The target variable was binary: 1 if the patient was readmitted within 30 days, 0 otherwise.

To begin, I explored the dataset to understand its structure and quality. I found the target classes were imbalanced: only 32.5% of patients were readmitted. This imbalance would later affect the model's ability to predict the minority class. Age stood out early as the feature most correlated with readmission, while other variables showed less distinction between classes. The dataset was clean, with a small number of outliers, and all columns were complete with no missing values.

In the preprocessing stage, I focused on preparing both the structured and unstructured data for modeling. I encoded categorical features like gender, diagnosis code, and medication type using one-hot encoding. For the free-text discharge notes, I manually extracted recurring clinical phrases (e.g., "good recovery trajectory," "continue current medication") and encoded them as binary flags. This allowed me to incorporate some of the text information without relying on more complex NLP methods, which would require more data to be effective. I also removed unnecessary columns like patient ID.

For modeling, I chose a Random Forest classifier due to its robustness and ease of interpretation. The first model achieved 65% accuracy. However, its recall for predicting readmitted patients was low (17%), which was expected given the class imbalance. After tuning hyperparameters, the model improved to 72% accuracy, and the ROC-AUC score increased to 0.83—suggesting strong overall classification performance. Still, recall for the minority class remained limited.

Feature importance confirmed earlier findings: age and length of stay were the most influential features, followed by number of previous admissions and some text-derived phrases like "continue current medication" and "blood pressure under control."

If I had more time and a larger dataset, I would take several next steps. First, I'd improve how I handle the discharge notes by using TF-IDF or word embeddings to better capture their meaning. Second, I'd address the class imbalance more directly using SMOTE or adjusting class weights. Finally, I'd experiment with ensemble methods like XGBoost or stacking to push recall and overall robustness further.