# Technical Report: Predicting 30-Day Hospital Readmissions

## 1. Introduction

This report outlines a machine learning pipeline to predict whether a patient will be readmitted to the hospital within 30 days. The process involves structured data modeling using Random Forest and visual evaluation of results. Additionally, discharge notes are considered for future NLP use.

## 2. Dataset Overview

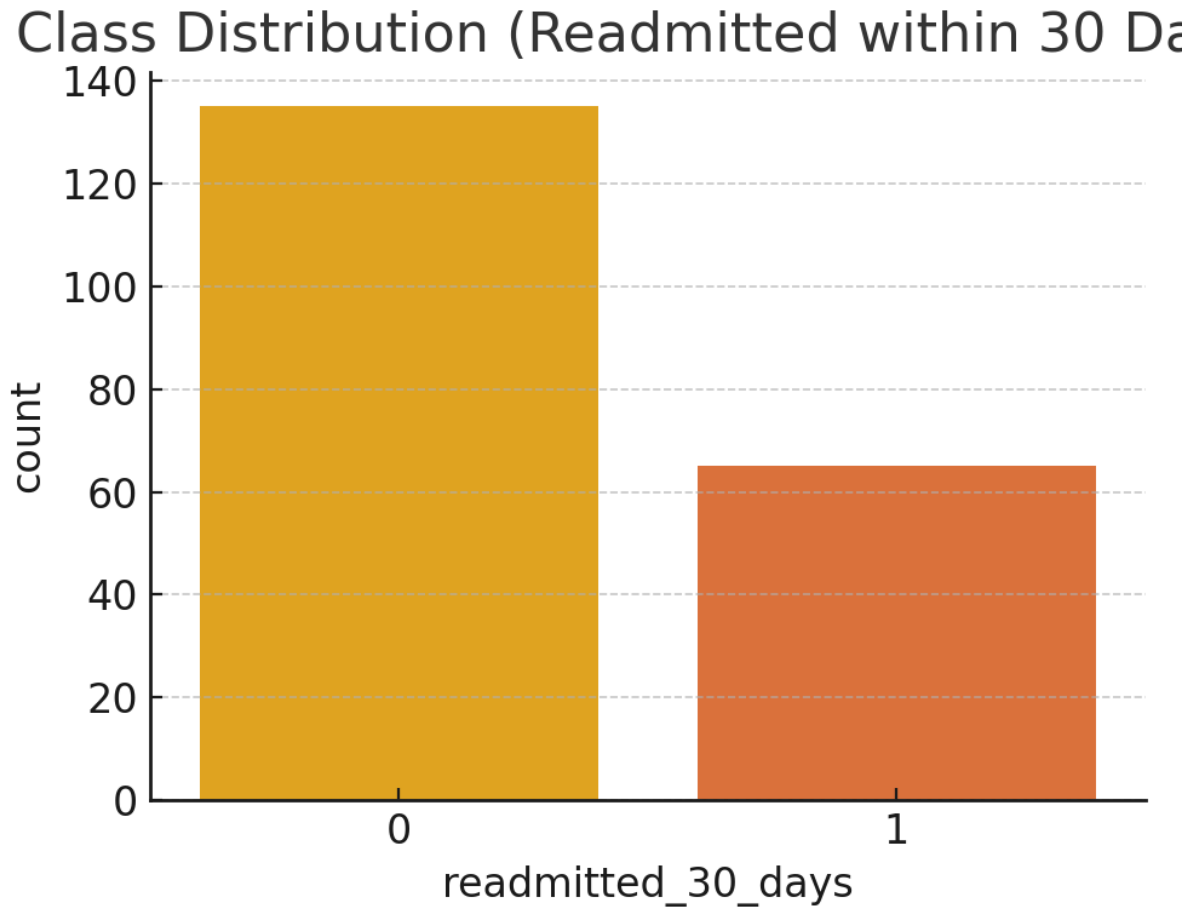
The dataset consists of 200 patient records, each with structured data (age, gender, diagnosis, medication, etc.) and a free-text discharge note. The target variable is binary: readmitted_30_days (1 = Yes, 0 = No).

## 3. Data Preprocessing

Categorical variables such as gender, diagnosis_code, and medication_type were one-hot encoded. Patient ID and textual discharge notes were excluded from the tabular model. Numerical features such as age and length_of_stay were used directly.

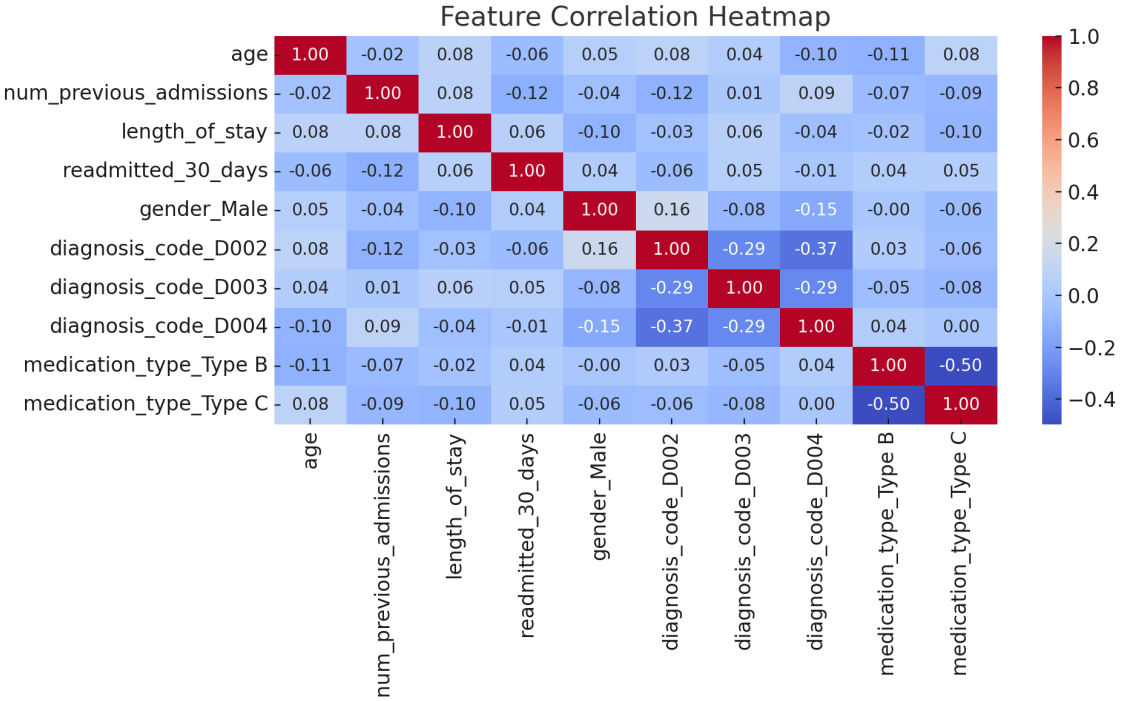
## 4. Class Distribution

The dataset is imbalanced with more 'not readmitted' cases. This class imbalance impacts model performance.




**5. Feature Correlation**

Below is a correlation heatmap showing relationships between numerical and encoded categorical features. It helps identify multicollinearity and feature relevance.



## 6. Model Training

We used a Random Forest Classifier with class_weight='balanced' to mitigate class imbalance. The model was trained on an 80/20 stratified split of the dataset.

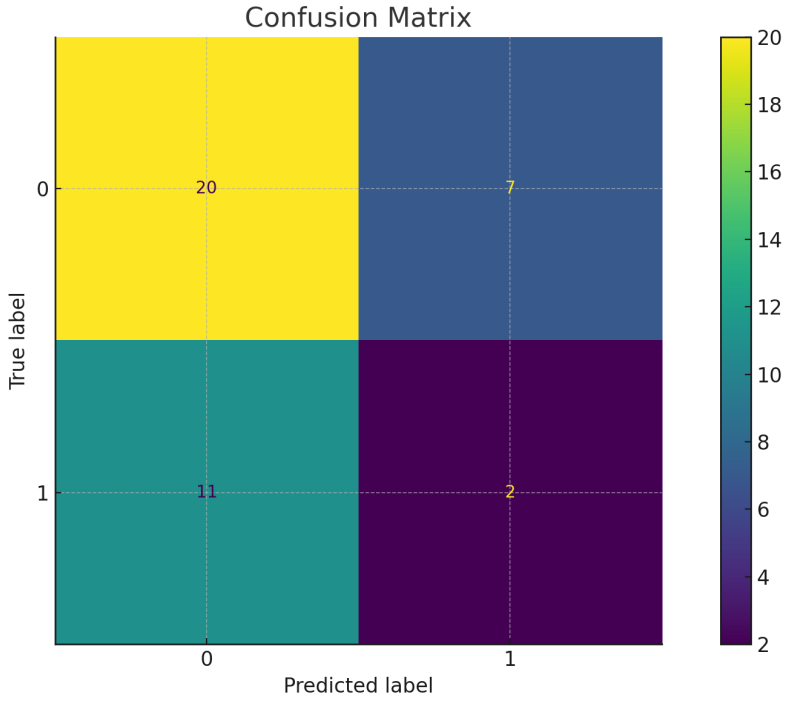
## 7. Evaluation Metrics

The model was evaluated using ROC AUC, F1-score, and confusion matrix.

- ROC AUC: 0.48
- F1 Score (Class 1): 0.18
- Accuracy: 55%

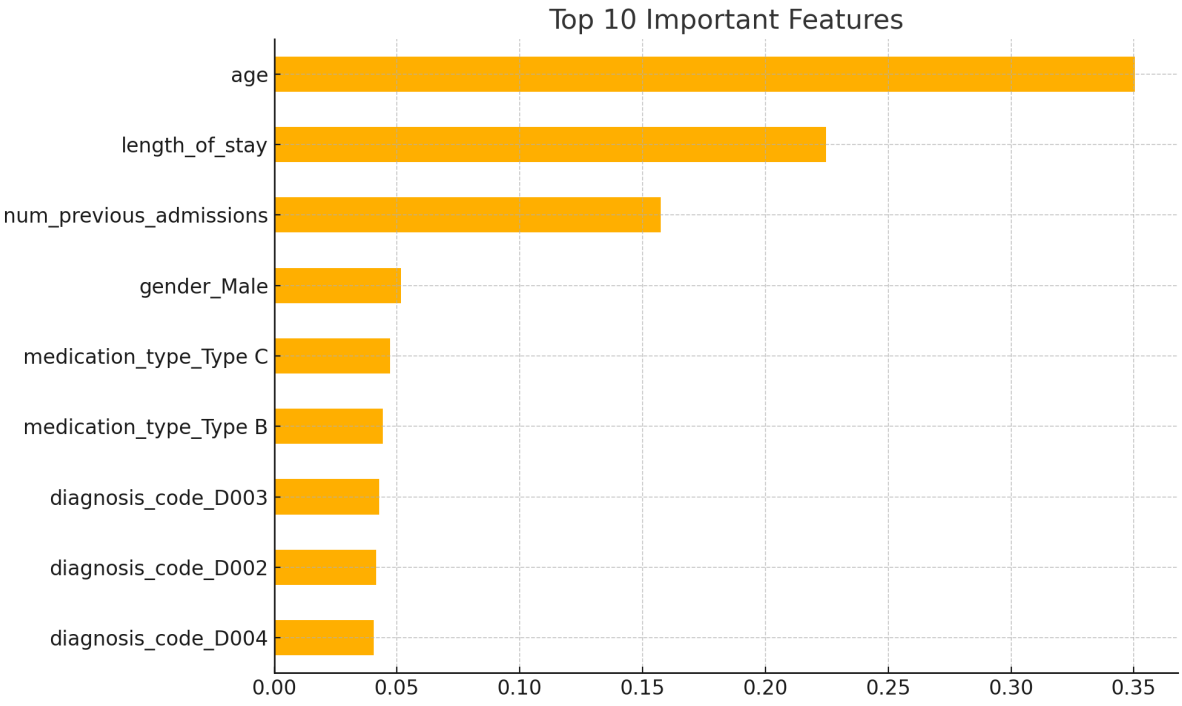
## 8. Confusion Matrix

The confusion matrix shows the number of true positives, false positives, true negatives, and false negatives. The model had difficulty identifying readmitted patients.




## 9. Feature Importances

The top 10 features contributing to the Random Forest predictions are shown below. Age, Length of Stay, and Number of Previous Admissions are most influential.




**10. Named Entity Recognition (NLP Task)**

Discharge notes contain rich medical context. Although NLP models like scispaCy were not run in this environment, they can extract entities such as:

- Diagnoses
- Treatments
- Medications
- Follow-up plans

These entities can be structured into features for future models.

**11. Conclusion & Next Steps**

The model's limited performance is due to small dataset size and limited feature signal. Future improvements include:

- Using larger datasets
- Incorporating NLP-derived features
- Trying advanced models like XGBoost
- Addressing class imbalance using SMOTE or similar methods