Predicting Hospital Readmission in Diabetic Patients

Machine Learning for 30-Day Readmission Risk Stratification

Galit Neufeld Kroszynski

Summary

Hospital readmission represent a critical challenge in modern healthcare, particularly for diabetic patients who face elevated risks of complications and return visits. Previous studies have identified several risk factors for readmission, such as prior hospitalizations and prolonged length of stay. However, traditional risk assessment methods often rely on limited clinical judgment and basic statistical models that fail to capture the complex interplay of factors influencing readmission. By leveraging advanced machine learning algorithms and extensive clinical data, we can transform readmission prevention from reactive to proactive care management.

101K

Hospitalizations Analyzed

Comprehensive dataset spanning 10 years across 130 hospitals

50+

Clinical Variables

Extensive feature set capturing patient complexity

66%

High-Risk Detection

Successfully identifies patients needing intervention

Therefore, this document presents an innovative comprehensive machine learning approach to predict 30-day readmission risk. Our predictive model analyzes over 50 clinical variables across more than 101,000 hospitalizations, identifying 66% of high-risk patients at discharge and providing explainable insights that empower clinical decision-making, enabling healthcare providers to deliver targeted interventions and improve patient outcomes. This data-driven approach represents a paradigm shift in how hospitals allocate resources and deliver post-discharge care.

The Problem: Diabetes and Readmission

Diabetes is a common chronic disease characterized by elevated blood glucose levels resulting from impaired insulin secretion or function. The disease affects millions of people worldwide and is associated with severe complications such as cardiovascular disease, kidney failure, eye and nerve damage, and impaired wound healing.

In the United States, approximately 10% of the population lives with diabetes, making it one of the leading causes of morbidity and mortality.

Hospital readmission of diabetic patients within 30 days of discharge poses a significant challenge to healthcare systems worldwide. The readmission rate among these patients stands at approximately 9% in the United States, significantly higher than in the general population.

Why Readmissions Matter

This phenomenon is problematic from three main perspectives:

From the patient's perspective

Readmission involves physical and emotional suffering, disruption to daily life, and increased risk of morbidity and mortality.

From the hospital's perspective

The average cost of readmission is approximately $15,000, and insurance companies are placing increasing pressure on hospitals to reduce readmission rates, sometimes through financial penalties.

From the healthcare system's perspective

Readmission places substantial strain on hospital capacity, compromise access to care for other patients, and indicate care delivery gaps that lead to preventable healthcare consumption among chronic patients

What is Currently Known

Traditional Approaches

- Retrospective statistical analysis
- Logistic regression models
- Limited clinical judgment
- Basic risk factor identification

Key Limitations

- Did not leverage advanced machine learning algorithms
- Failed to capture complex factor interactions
- Lacked practical prediction tools
- Insufficient for real-time decision support

Previous studies have identified several risk factors for readmission, including: prior hospitalizations, prolonged length of stay, high number of medications, multiple comorbidities, and advanced age. However, most studies focused on retrospective statistical analysis using traditional methods such as logistic regression, and did not leverage the potential of advanced machine learning algorithms for prediction and classification.

Project Objectives

This project aims to fill the existing gap by developing an advanced prediction model, based on machine learning algorithms, to identify diabetic patients at high risk for readmission within 30 days of discharge. The goal is to provide medical teams with a practical tool that assists in focusing on patients requiring special attention and targeted interventions.

1

Build a predictive model

Identify high-risk patients for early intervention

2

Reduce readmission rates

Improve patient outcomes through targeted care

3

Lower healthcare costs

Optimize resource allocation and reduce financial burden

Data Overview

Time Period

10 years (January 1999 - December 2008)

Hospitals

130 facilities across the United States

Hospitalizations

101766 Unique patient encounters

Data Source

Health Facts Database from Cerner Corporation, UCI Machine Learning Repository

Data Files

1. **Diabetic_data.csv** - the main table file contains approximately 101,760 unique hospitalizations from 130 hospitals in the United States.
2. **IDS_mapping.csv** - file contains all the code mappings.

Inclusion Criteria

- Inpatient admission with diabetes diagnosis or consuming diabetes medications
- Length of stay: 1-14 days
- Laboratory tests and medications administered

Target Variable

Outcome: 30-day hospital readmission

Class 1 (Readmitted <30 days): 11% of cases

Class 0 (Not readmitted <30 days): 89% of cases

Imbalance Ratio: 1:8 (minority:majority)

Explanatory Variables

The dataset contains 47 original features:

Demographics

Age, gender, race

Admission/Discharge

Type, source, disposition, time in hospital, payer code

Clinical Complexity

Lab procedures, procedures, medications, diagnoses, diagnosis codes

Medical History

Outpatient/emergency/inpatient visits (prior year)

Glycemic Control

Max glucose serum, A1C test result

Diabetes Medications

23 specific medications with dosage changes

Data Preprocessing Steps

Step 1: Remove Invalid Records

Removed 2,447 encounters 2.4%) where readmission impossible: Expired patients, Hospice discharges, stayed at admission

Final Dataset: 99,319 valid encounters

Step 2: Handle Missing Values

Weight: Removed (97% missing)

A1C test: Retained as binary flag no_A1C_test (83% missing - indicates care quality gaps)

Small-scale missing (<0.1%): Imputed to majority class

Race, gender, medical specialty: Missing treated as "Unknown" category

Step 3: Consolidate Categorical Variables

Clinical Domain Knowledge Integration: All categorical consolidations guided by clinical expertise, not purely data-driven. Consulted medical literature and clinical guidelines to ensure clinically meaningful groupings.

Variable Consolidation

Clinical rationale for categorical variable groupings:

Variable	Original	Consolidated	Clinical Rationale
Age	10 brackets	4 groups	Young 0-40, Middle 40-60, Older 60-80, Elderly 80+. Reflects diabetes onset patterns and complication risks
Medical Specialty	73 specialties	17 groups + Missing	Grouped by care delivery model and diabetes management role
Payer Code	18+ codes	6 categories	Groups by socioeconomic status and access patterns
Admission Type	8 types	4 groups	Emergency, Urgent, Elective, Not_Available. Trauma Center (7) merged into Emergency. Fixed "Newborn" coding errors
Discharge Disposition	28 codes	5 groups	Home, Transfer_to_Facility, Hospice, Left_AMA, Unknown. Critical predictor - care transitions introduce errors
Admission Source	26 codes	5 groups	Emergency, Referral, Transfer_from_Facility, Readmission_HHA, Unknown. Reflects patient acuity. Fixed birth-related coding errors
ICD-9 Diagnoses	1,000+ codes	11 clinical categories	Grouped by diabetes-specific relevance and readmission risk

Data Exploratory Strategy

We created initial reports using ydata-profiling to identify data behavior. Also, we used the classic descriptive measurements of Min, Max, Mean to observe significant trends.

This revealed substantial deviations from normality across most numerical features. Variables such as **number_emergency**, **number_outpatient**, and **num_procedures** exhibited extreme right-skewed distributions with high concentrations of zero values and exponential decay patterns

. In contrast, **num_lab_procedures** and **number_diagnoses** displayed more symmetric, near-normal distributions. These distributional characteristics necessitate transformation techniques (logarithmic) or non-parametric modeling approaches for subsequent statistical analysis.

We also conducted Statistical Tests: **Mann-Whitney U** (numerical-target), **Chi-Square** (categorical-target), **Cramér's V** and **Cohen's d** (effect size) to explore relationships of different variables. While most features demonstrated statistical significance due to the large sample size, effect size analysis revealed that the majority lack practical predictive value, with only prior hospitalization history showing meaningful clinical impact.

These findings suggest that feature engineering and the creation of composite variables may be necessary to develop clinically relevant predictors beyond the existing feature set.

Top Features (by effect size):

1

number_inpatient:

Cohen's d = 0.536 (Medium)

2

number_emergency:

Cohen's d = 0.192 (Negligible)

3

number_diagnoses:

Cohen's d = 0.1659 (Negligible)

4

discharge_grouped:

Cramér's V = 0.085 (Negligible)

5

insulin:

Cramér's V = 0.044 (Negligible)

6

diag_3_category:

Cramér's V = 0.041 (Negligible)

Data cleansing

Outlier management was implemented through two complementary strategies tailored to different distributional characteristics. First, **log transformation (log1p)** was applied to features exhibiting severe right-skewness and zero-inflation, including **number_inpatient**, **number_outpatient**, **number_emergency**, **num_procedures**, **num_medications**, and derived features such as **total_prior_visits** and ratio features (**meds_per_day**, **labs_per_day**, **procedures_per_day**). The log1p transformation—log(x+1)—was specifically chosen to handle zero values gracefully while compressing extreme values and normalizing distributions, thereby reducing the disproportionate influence of outliers.

Second, **quantile-based capping (Winsorization)** was applied post-split to features with long-tailed distributions but without zero-inflation: **time_in_hospital**, **num_lab_procedures**, and **number_diagnoses**. For these features, values below the 1st percentile and above the 99th percentile were clipped to their respective thresholds. Critically, capping boundaries were learned exclusively from the training set to prevent data leakage and then consistently applied to validation and test sets.

Feature Engineering

Created 20 new features:

Intensity Metrics

Normalized by length of stay:

- meds_per_day
- labs_per_day
- procedures_per_day
- total_prior_visits

Risk Indicator Flags

- had_prior_inpatient
- high_complexity
- no_A1C_test

Medication Pattern Features

- num_diabetes_meds_active
- insulin_changed
- diabetes_med_changed

Composite Risk Scores

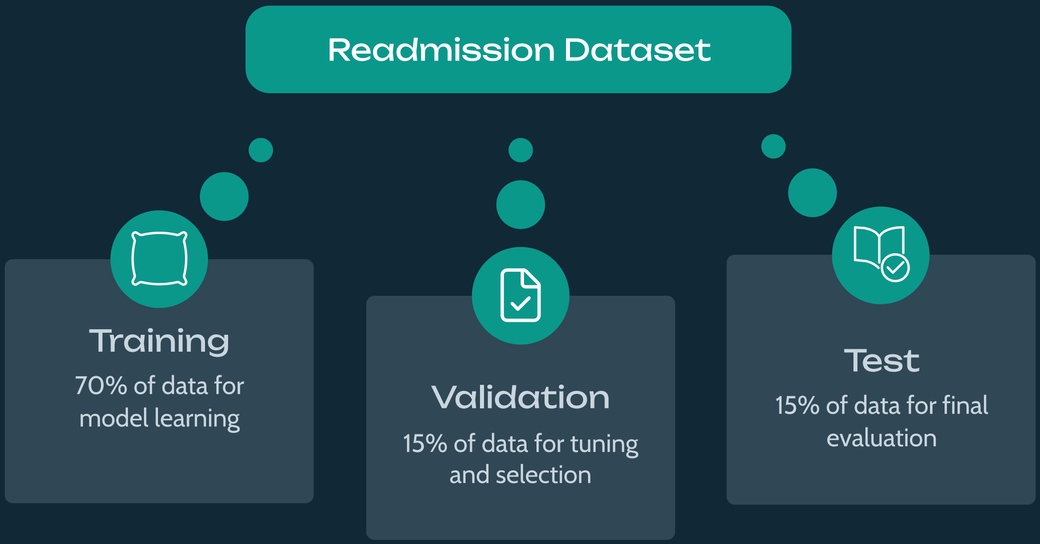
- high_risk_patient
- is_hidden_risk_profile

Interaction Features

- A1C_NM_AND_med_change
- no_prior_inp_AND_med_change
- short_stay_AND_med_change.

Train-Validation-Test Data Split

The whole data was divided into three separate sets: Training, Validation, and Test, with proportions of 70%, 15%, and 15%, respectively. The split was performed in two stages: first, 70% of the data was separated for the training set and 30% for a temporary set, then the temporary set was split equally between the validation and test sets (each comprising 15% of the original data). To ensure result reliability, the split was performed with stratification - maintaining the same class distribution (readmitted/not readmitted) across all three sets, which is critical given the significant class imbalance in the dataset (approximately 11% readmission rate).



Readmission Dataset

Training

70% of data for model learning

Validation

15% of data for tuning and selection

Test

15% of data for final evaluation

Handling Class Imbalance

Challenge: 11% readmitted vs. 89% not readmitted.

Tested three techniques:

SMOTE

PR-AUC = 0.138

Recall = 0.08

RandomOverSampler

PR-AUC = 0.192

Recall = 0.56

RandomUnderSampler

PR-AUC = 0.191

Recall = 055

Final Training Set: 61,611 Class 0 + 61,611 Class 1 = 123222 samples (50:50 balance)

Encoding, and Scaling

- **Encoding:** One-hot encoding → expanded from 69 to 117 features
- **Scaling:** StandardScaler for numerical features (mean=0, std=1)

Feature Selection

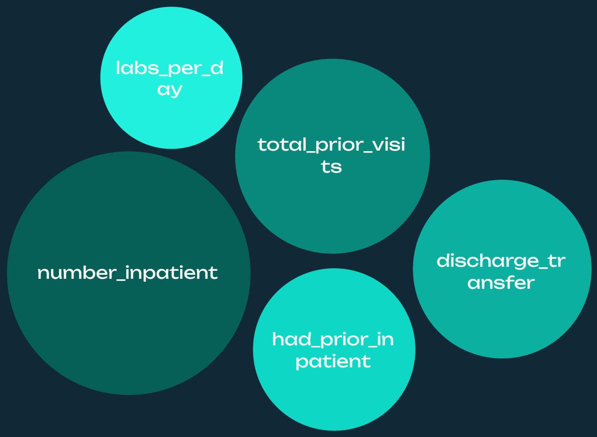
Method: Random Forest feature importance analysis

Decision: Retained 50 features with importance ≥0.005 (57% reduction from 117)

Validation: Model performance on 117 vs. 50 features showed negligible difference (PR-AUC <0.005)

Top 10 Features:

number_inpatient	0.102097
total_prior_visits	0.081011
had_prior_inpatient	0.076629
discharge_grouped_Transfer_to_Facility	0.064369
labs_per_day	0.036979
meds_per_day	0.036319
num_medications	0.035088
num_lab_procedures	0.033975
number_diagnoses	0.031699
high_risk_patient	0.030909



Model Training & Comparison

Objective: Maximize recall to identify high-risk patients while maintaining acceptable precision.

Models Evaluated: Decision Tree, Random Forest, AdaBoost, Gradient Boosting, XGBoost, LightGBM. We also evaluated a CatBoost model which is a

Best Performance (Validation Set)

Model	PR-AUC	Recall	Precision
CatBoost ✓	0.204	0.6	0.178
Decision Tree	0.177	0.59	0.15
Adaboost	0.194	0.57	0.17

Hyperparameter Tuning

We employed Optuna with Bayesian optimization (TPE sampler) to tune CatBoost hyperparameters across 40 trials, targeting maximum recall. MedianPruner enabled early stopping of underperforming trials, reducing training time by ~30%. This intelligent approach outperformed random search by probabilistically modeling the objective function and concentrating resources on promising parameter regions. Recall score improved to **0.6763**.

Threshold Optimization

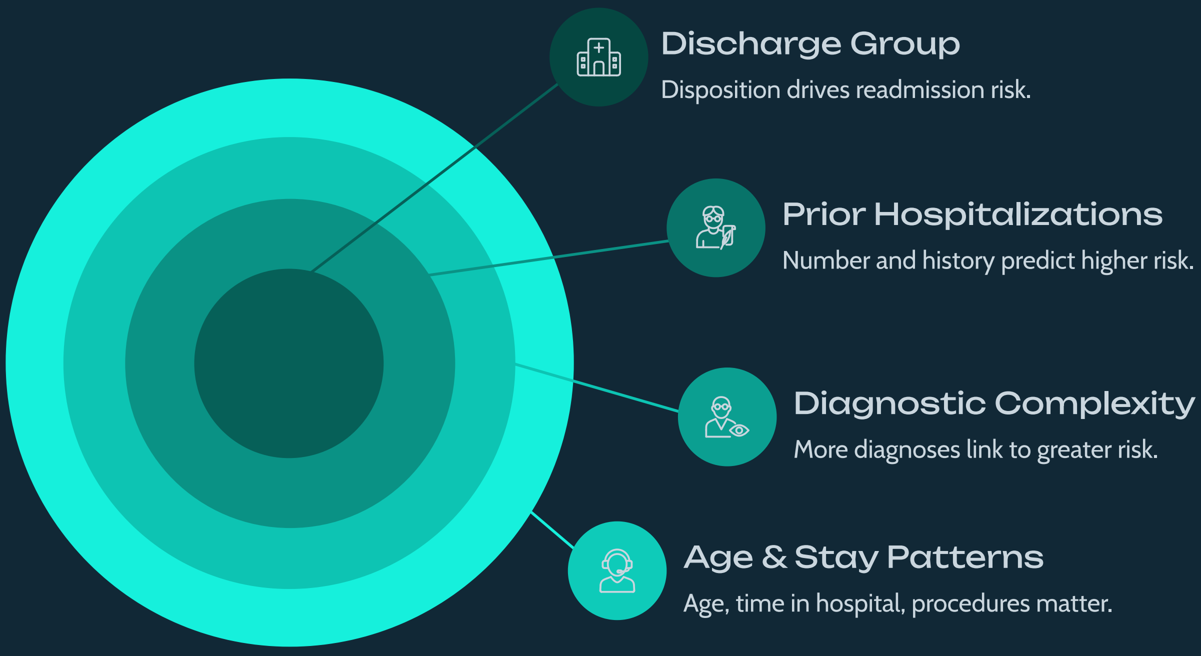
We conducted a specific analysis to determine the optimal prediction threshold for the final model. We compared the default 0.5 threshold against an F1-optimized threshold (0.5102), which was calculated on the validation set. While the F1-optimized threshold offered a minor increase in precision (17.0% vs. 15.9%), it came at a significant and unacceptable cost to our primary business objective. Using this higher threshold caused the model's test set recall to drop dramatically from **0.667** to **0.568**, resulting in 10% more high-risk patients being missed. Given that the goal is to maximize patient identification, we concluded that the default **0.5 threshold** is the superior choice, as it provides the best recall performance.

Key Predictive Features - SHAP Analysis

SHAP (SHapley Additive exPlanations) analysis revealed the most influential features driving readmission predictions in the CatBoost model:

The top predictor was **discharge_grouped** (discharge disposition), showing wide variability in impact across different discharge destinations, suggesting that post-discharge care arrangements critically affect readmission risk. **Number_inpatient** (prior inpatient visits) emerged as the second most important feature, with a clear positive correlation: patients with higher prior hospitalization counts (red/pink points) consistently showed elevated SHAP values, indicating substantially increased readmission risk.

This pattern was reinforced by **had_prior_inpatient** and **total_prior_visits**, confirming that hospitalization history serves as a powerful prognostic indicator. **Number_diagnoses** demonstrated a complex non-linear relationship, where increased diagnostic complexity (higher values) correlated with higher readmission risk. **Age_ordinal** showed the expected clinical pattern, with older patients (red points) exhibiting positive SHAP values and younger patients (blue points) showing negative values. Notably, **time_in_hospital** and **procedures_per_day** displayed bidirectional distributions, reflecting U-shaped relationships where both extremely short and extremely long hospitalizations, as well as very low or very high procedure rates, increased readmission risk. Interestingly, medication-related features (**insulin**, **num_diabetes_meds_active**) showed relatively lower importance compared to clinical history and demographic factors, suggesting that patient baseline risk factors outweigh acute treatment decisions in predicting 30-day readmissions.



Discharge Group

Disposition drives readmission risk.

Prior Hospitalizations

Number and history predict higher risk.

Diagnostic Complexity

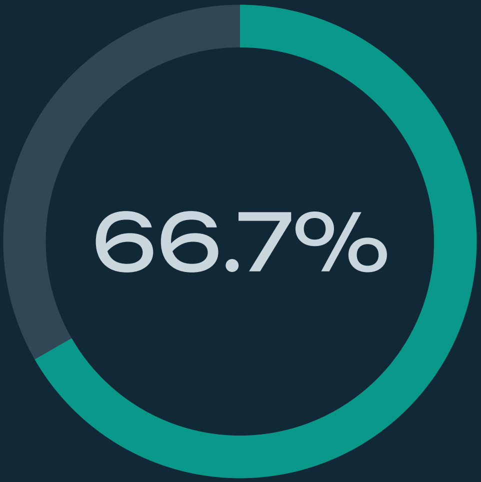
More diagnoses link to greater risk.

Age & Stay Patterns

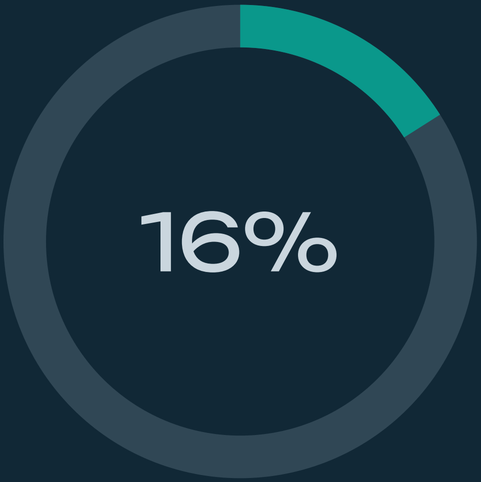
Age, time in hospital, procedures matter.

Final Model Evaluation

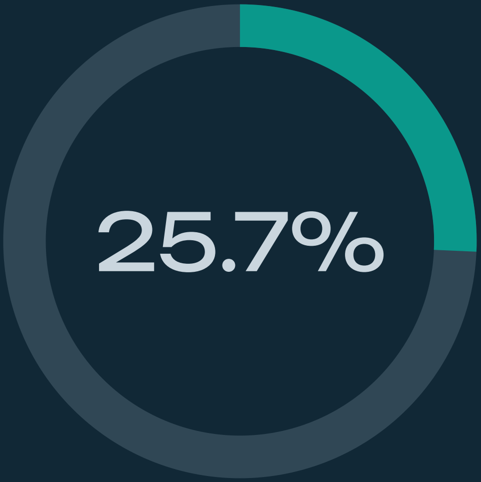
Final Performance (Test Set)



Recall

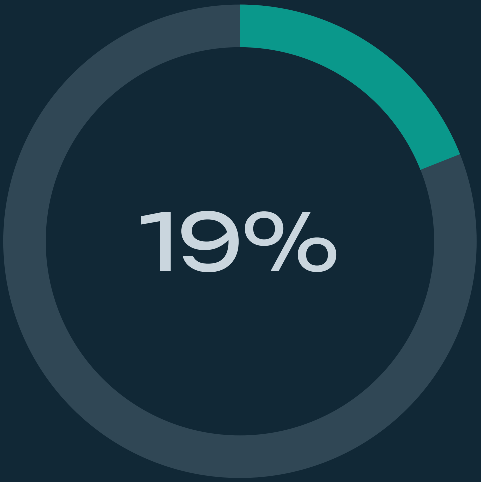


Precision



F1-Score

Harmonic mean



PR-AUC

Strong performance for imbalanced data

Model Robustness (Validation vs. Test)

The final model demonstrates strong robustness and generalization. During the tuning phase, the best-performing model achieved a **validation recall of 0.676**. When this model was applied to the unseen test set, it achieved a nearly identical **test recall of 0.667**. This negligible gap of less than 1% confirms that the model is not overfit and has successfully learned the true underlying patterns in the data.

How to use this model?
Clinical Workflow Integration

Step 1: Data Collection

During Hospitalization - All required features captured automatically in Electronic Health Record (EHR). No manual data entry required.

Step 2: Risk Prediction

At Discharge - Trigger: When discharge order is entered (24-48 hours before discharge). Model Runs: Extracts 49 features from EHR, generates prediction in <5 seconds. Output: Risk score (0-100% probability), Risk tier (Red/Yellow/Green), SHAP explanation (top 5 features driving the prediction).

Step 3: Clinical Review

Alert to Case Manager: Automated notification in EHR. Review Risk Factors: Case manager reviews SHAP explanation. Assign Interventions based on risk tier.

Step 4: Post-Discharge Monitoring

Track actual readmissions for 30 days. Compare predictions to outcomes. Retrain model annually with new data.

Intervention Tiers

High-Risk (Red Tier, 52.3%)

- Intensive care coordination (dedicated coordinator for 30 days)
- Home health nurse visit within 48-72 hours
- Primary care appointment within 7 days

Medium-Risk (Yellow Tier, 51.48%- 52.3%)

- Automated follow-up calls (days 3 and 7)
- Diabetes self-management education materials
- Doctor's appointment within 14 days

Low-Risk (Green Tier, <52.3%)

- Standard discharge protocol

Appendix: Project Code Structure

1

Data Preparation (01_Data_Preparation)

This first notebook includes loading raw data, filtering records irrelevant to prediction (such as deceased patients), handling missing values, and performing basic feature engineering. The output of this stage is a clean dataset ready for analysis.

2

Exploratory Data Analysis (02_Exploratory_Data_Analysis)

The second notebook is dedicated to a deep understanding of the data. In this stage, statistical tests (such as Mann-Whitney U for continuous variables and Chi-Square for categorical variables) were performed to identify significant relationships between clinical features and the target variable (readmission).

3

Modeling & Evaluation (03_Modeling_and_Evaluation)

The third notebook contains the core Machine Learning work. This stage includes advanced feature engineering (medical complexity, interactions), handling imbalanced data using undersampling, and training and comparing various models. Finally, this notebook performs optimization of the selected model (CatBoost) and explainability analysis using SHAP values.