VISHAL CHAUDHARY · DATA ANALYST · DUBLIN, IRELAND

Healthcare Machine Learning · Academic Project

Diabetes Readmission Prediction

Predicting 30-day hospital readmissions for diabetic patients using interpretable machine learning on 101,763 patient records.

101,763	82%	0.65	47
Patient Records	Non-Readmission Recall	ROC-AUC Score	Input Features

TOOLS & TECHNOLOGIES

Python	Scikit-learn	Logistic Regression	SMOTE	Pandas	Statsmodels
IBM Cognos	Matplotlib				

Email vishal.ch1401@gmail.com	LinkedIn linkedin.com/in/vishal111	GitHub github.com/chaudhary521	Location Dublin, Ireland

PROBLEM STATEMENT

20% of diabetic patients in the US are readmitted within 30 days of discharge, driving up healthcare costs and signalling gaps in care quality. The challenge was to build a predictive model that flags high-risk patients before discharge so clinical teams can intervene early and reduce unnecessary readmissions.

DATASET

UCI Diabetes 130-US Hospitals dataset: 101,763 patient visits across 130 US hospitals spanning 1999 to 2008. The dataset contains 47 features including medication types, number of diagnoses, lab test results, length of stay, and patient demographic information. Data required extensive cleaning for missing values and inconsistent coding.

APPROACH & METHODOLOGY

Logistic Regression was chosen as the primary model for its clinical interpretability — understanding why a patient is flagged is as important as the prediction itself. The pipeline covered full EDA, removal of near-zero-variance features and patient identifiers, binary target encoding (readmitted within 30 days = 1), SMOTE oversampling to address class imbalance, and iterative feature elimination using statistical significance ($p < 0.05$). Model outputs were integrated into an IBM Cognos Analytics dashboard for non-technical clinical stakeholders.

KEY TECHNICAL HIGHLIGHTS

› Applied SMOTE oversampling to correct severe class imbalance between readmitted and non-readmitted patients.

› Feature selection via statistical significance testing ($p < 0.05$) retained only clinically meaningful predictors.

› Achieved ROC-AUC of 0.65 and correctly identified 82% of non-readmission cases.

› Top predictors: number of inpatient visits, time in hospital, number of diagnoses, and lab procedures count.

› Deployed model outputs in an IBM Cognos Analytics dashboard accessible to non-technical clinical staff.

› Completed as academic research under the MSc Data Analytics programme at NCI Dublin.

KEY INSIGHTS & RESULTS

The final model achieved 61.73% overall accuracy and ROC-AUC of 0.65. Extended hospital stays and high prior inpatient visit counts were the strongest readmission risk signals. The model correctly identified 82% of non-readmissions, making it practically useful for triage prioritisation. Draw prediction (in this context, borderline cases) remained the most uncertain class.

BUSINESS IMPACT

Enables clinical teams to identify high-risk diabetic patients before discharge, allowing for targeted follow-up protocols, care coordination, and medication reviews. Early identification has the potential to reduce 30-day readmission rates and the associated financial and patient welfare costs. The Cognos dashboard makes insights accessible without requiring clinical staff to interact directly with the model.
