Hospital Readmission Prediction Documentation

# Introduction

Hospital readmission refers to the phenomenon where a patient is admitted to the hospital again within a certain period after their initial discharge. It is an important metric in healthcare management as it reflects the quality of care provided during the initial hospitalization and can indicate potential gaps in care continuity. Predicting hospital readmissions allows healthcare providers to identify patients at risk and intervene early to prevent unnecessary hospitalizations, improve patient outcomes, and reduce healthcare costs. Utilizing machine learning (ML) for readmission prediction enables healthcare institutions to leverage patient data to develop accurate predictive models that can assist in early identification and intervention.

<https://physionet.org/content/mimiciv/2.2/>

# Data Source

The dataset used for this analysis is sourced from the MIMIC-IV database version 2.2, available on PhysioNet. The dataset contains de-identified electronic health records from patients admitted to the Beth Israel Deaconess Medical Center from 2011 to 2019. After downloading the dataset, it is loaded into a Redshift database for further analysis and processing.

# Preprocessing Steps

1. *Merge Admission Data with Diagnosis Data*: Admission records are merged with diagnosis data to associate each admission with the corresponding medical conditions diagnosed during the stay.

2*. Merge Admission Data with Chart Events*: Admission data is further enriched by merging it with chart events data, providing detailed clinical observations recorded during the patient's stay.

3*. Group Patient Label Data with Hospital Admissions*: Patient label data, such as readmission status, is grouped with admission records to create a labeled dataset for predictive modeling.

4. *Recategorize Admission Type and Race*: Admission types and race categories are recategorized into broader groups for simplification and improved model interpretability.

# Rules Applied

1. Recategorize Admission Type: "EW EMER.", "URGENT", "DIRECT EMER." are categorized as "Emergency".
2. "OBSERVATION ADMIT", "EU OBSERVATION", "DIRECT OBSERVATION", "AMBULATORY OBSERVATION" are categorized as "Observation".
3. "SURGICAL SAME DAY ADMISSION" is categorized as "Surgical Same Day Admission".
4. "ELECTIVE" is categorized as "Elective".
5. Recategorize Race: Various subgroups are grouped under broader categories such as "ASIAN", "BLACK/AFRICAN", "HISPANIC OR LATINO", etc.
6. Mapping of ICD Codes to Categories:

|  |  |
| --- | --- |
| **ICD Range** | **Chapter Block Title** |
| 001–139 | Certain infectious and parasitic diseases |
| 140–239 | Neoplasms |
| 240–279 | Endocrine, nutritional and metabolic diseases and immunity disorders |
| 280–289 | Diseases of the blood and blood-forming organs |
| 290–319 | Mental disorders |
| 320–389 | Diseases of the nervous system and sense organs |
| 390–459 | Diseases of the circulatory system |
| 460–519 | Diseases of the respiratory system |
| 520–579 | Diseases of the digestive system |
| 580–629 | Diseases of the genitourinary system |
| 630–679 | Complications of pregnancy, childbirth, and the puerperium |
| 680–709 | Diseases of the skin and subcutaneous tissue |
| 710–739 | Diseases of the musculoskeletal system and connective tissue |
| 740–759 | Congenital anomalies |
| 760–779 | Certain conditions originating in the perinatal period |
| 780–799 | Symptoms, signs, and ill-defined conditions |
| 800–999 | Injury and poisoning |
| E and V | External causes of injury and supplemental classification (E codes); (V codes) |
| XXI | Factors influencing health status and contact with health services |
| XXII | Codes for special purposes |

# Exploratory Data Analysis (EDA)

The distribution of diagnoses across different categories is analyzed to understand the prevalence of various medical conditions among the patient population. Categories are derived based on the mapping of ICD codes.

# Model Building

Random Forest Classifier is used for building predictive models to predict readmissions at different time intervals (30 days, 60 days, and 365 days).

# Model Evaluation

The performance of the models is evaluated using standard classification metrics such as precision, recall, F1-score, and accuracy.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Time Interval** | **Precision (0)** | **Recall (0)** | **F1-score (0)** | **Precision (1)** | **Recall (1)** | **F1-score (1)** | **Accuracy** |
| 30 days | 0.9 | 0.99 | 0.94 | 0.97 | 0.74 | 0.84 | 0.92 |
| 60 days | 0.91 | 0.97 | 0.94 | 0.95 | 0.87 | 0.91 | 0.92 |
| 365 days | 0.91 | 0.99 | 0.95 | 0.98 | 0.76 | 0.85 | 0.93 |

# Deployment

The trained models are deployed using pickle files, facilitating easy integration into production systems for real-time prediction of readmission risks. This deployment approach enables proactive care management and resource allocation, leading to improved patient outcomes and healthcare efficiency.

# Files Used

1. Readmission\_Data\_Exploration\_Retrain:

* This file is used for training and generating the model, along with preprocessing steps.
* It likely contains code for loading the dataset, exploring its structure, handling missing values, performing feature engineering, and preparing the data for model training.
* Additionally, it may include code for training machine learning models, evaluating their performance, and tuning hyperparameters.

2. EDA\_Readmissions.ipynb:

* This notebook is dedicated to exploratory data analysis (EDA) of the Mimic dataset used for readmission prediction.
* It involves analyzing the distribution of features, identifying patterns and correlations, visualizing data using plots and charts, and gaining insights into the dataset's characteristics.
* EDA helps in understanding the dataset better and informing subsequent preprocessing and modeling steps.

3. ModelBuilding.ipynb:

* This notebook focuses solely on model building.
* It may include code for loading preprocessed data, splitting it into training and testing sets, and building machine learning models.
* Since preprocessing steps can be time-consuming, this notebook might have been created to save time by skipping those steps and directly focusing on training models.

4. Readmission\_Model\_Random\_100.ipynb:

* Like the previous notebook, this file is used for model building.
* However, it specifically mentions using a subset of 100 patient IDs for modeling purposes.
* It's possible that this subset was chosen for quicker experimentation or testing of model algorithms and parameters.