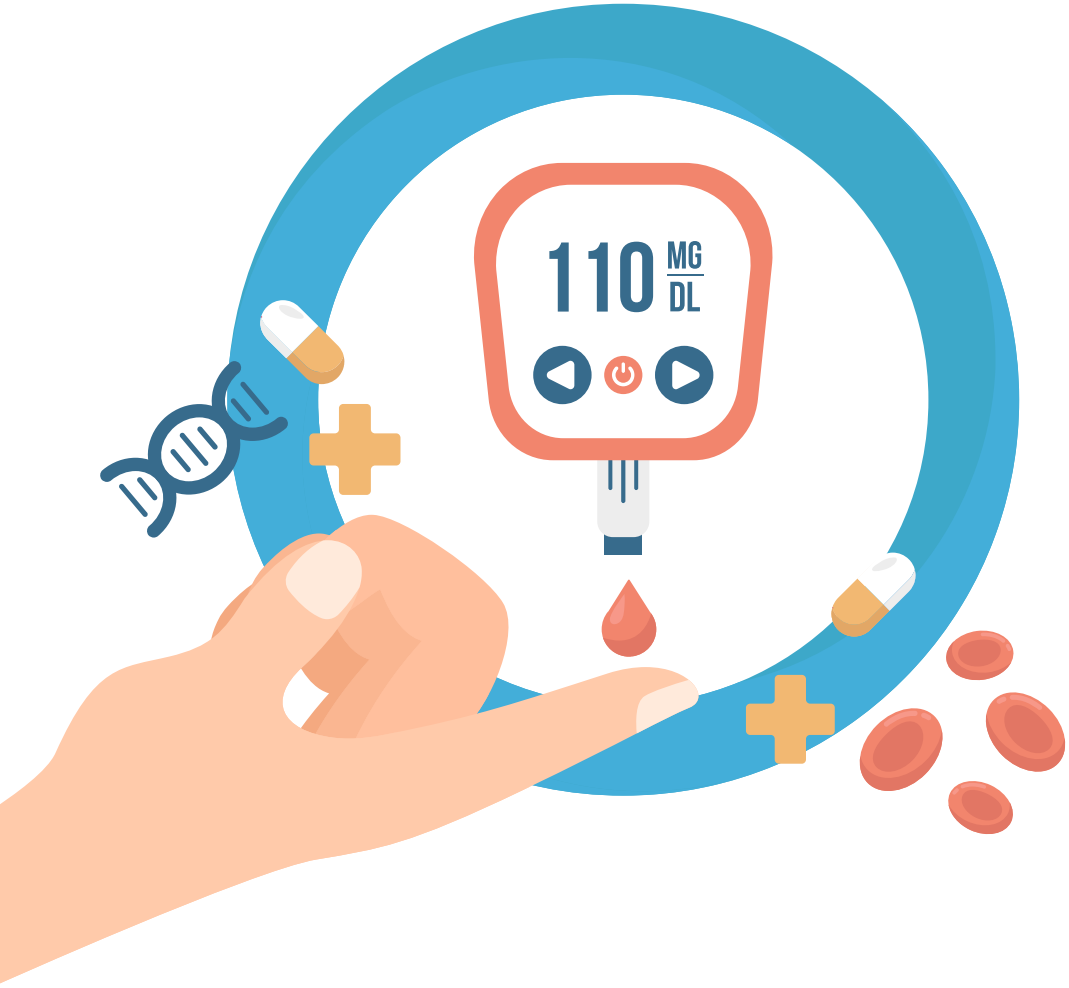
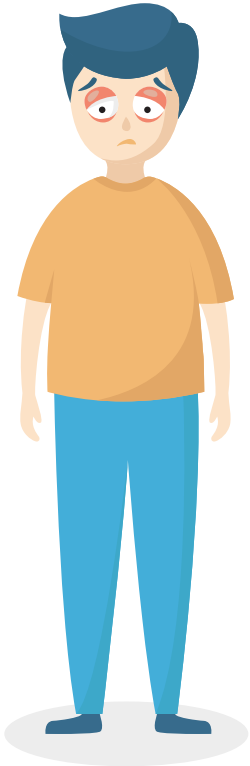


Diabetes Data - Analysis and Patient Readmission Prediction

Introduction



- Diabetes is a chronic disease that occurs when the body cannot effectively use the insulin it produces or does not produce enough insulin. This leads to an increased concentration of glucose in the blood, which can lead to long-term damage to the body and failure of various organs and tissues.

- According to the World Health Organization, an estimated 422 million adults were living with diabetes in 2014, compared to 108 million in 1980. The global prevalence of diabetes among adults over 18 years of age has risen from 4.7% in 1980 to 8.5% in 2014.

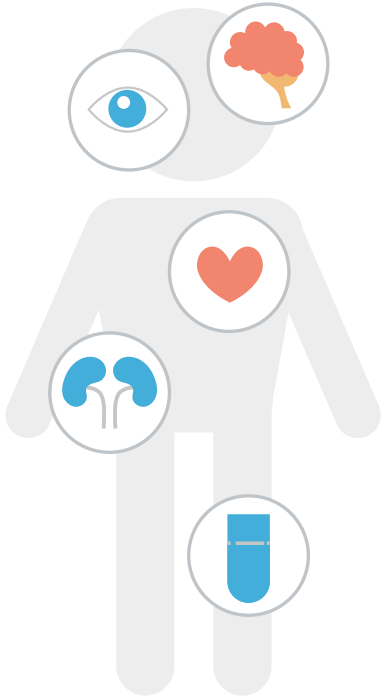
- One of the major concerns for diabetic patients is hospital readmission. Readmission not only leads to increased healthcare costs but also indicates potential deterioration of the patient's health. Predicting readmission can lead to improved patient care and management.

Diabetes infographic

Main types

Type 1, Type 2 and Gestational Diabetes

Organs affected



Prevention

Balanced diet, Regular exercise, Healthy weight

Symptoms

Increased thirst, frequent urination, fatigue, blurred vision, numbness, unhealed sores, weight loss.

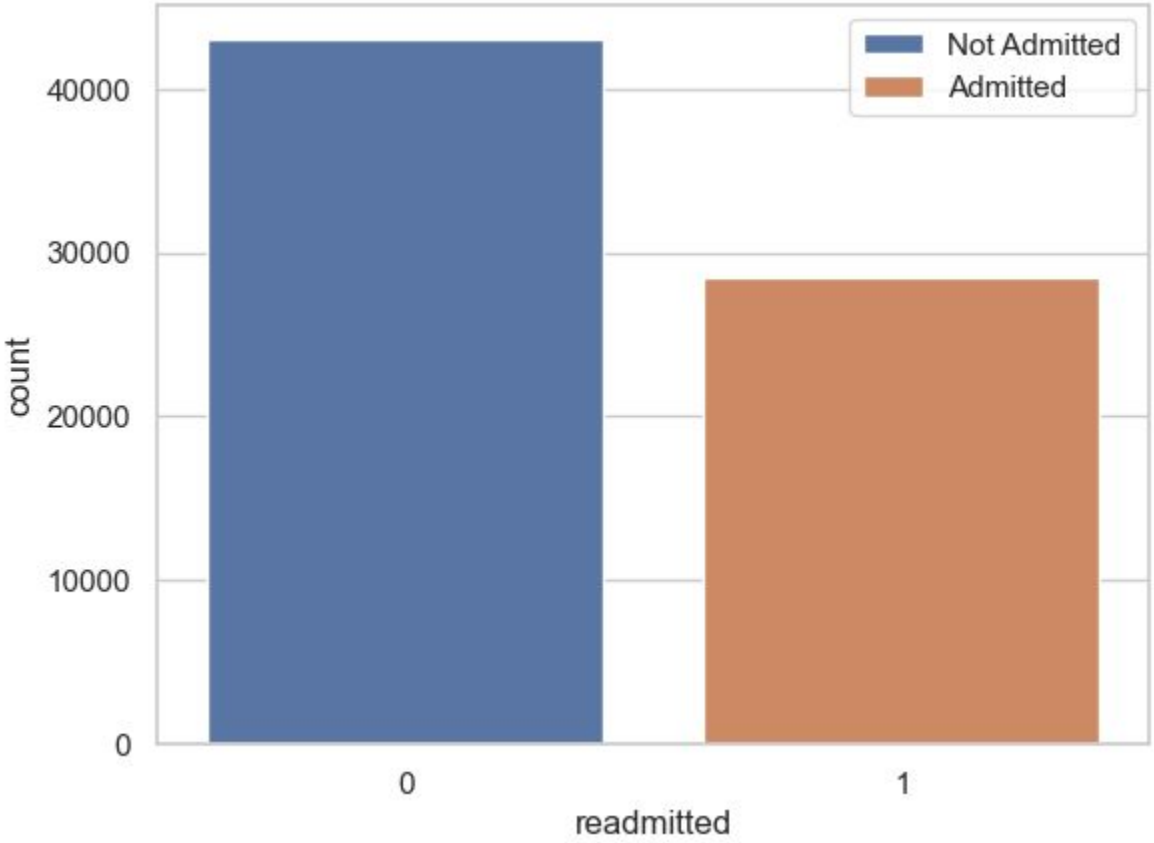
Treatment

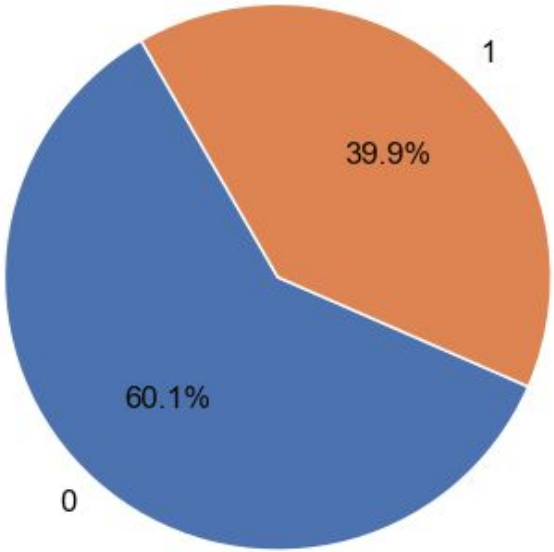
Insulin, Oral medications, Transplantation

Objective

- We conducted a comprehensive analysis of the diabetes dataset, examining various features such as patient demographics, diagnoses, number of visits, medication, etc.
- Our analysis identified correlations among the features in the dataset. Understanding these relationships provided insights into factors that influence patient readmission.
- Using machine learning models, we predicted the likelihood of patient readmission. This predictive tool will assist healthcare providers in making informed decisions and taking preventive measures to reduce readmissions.

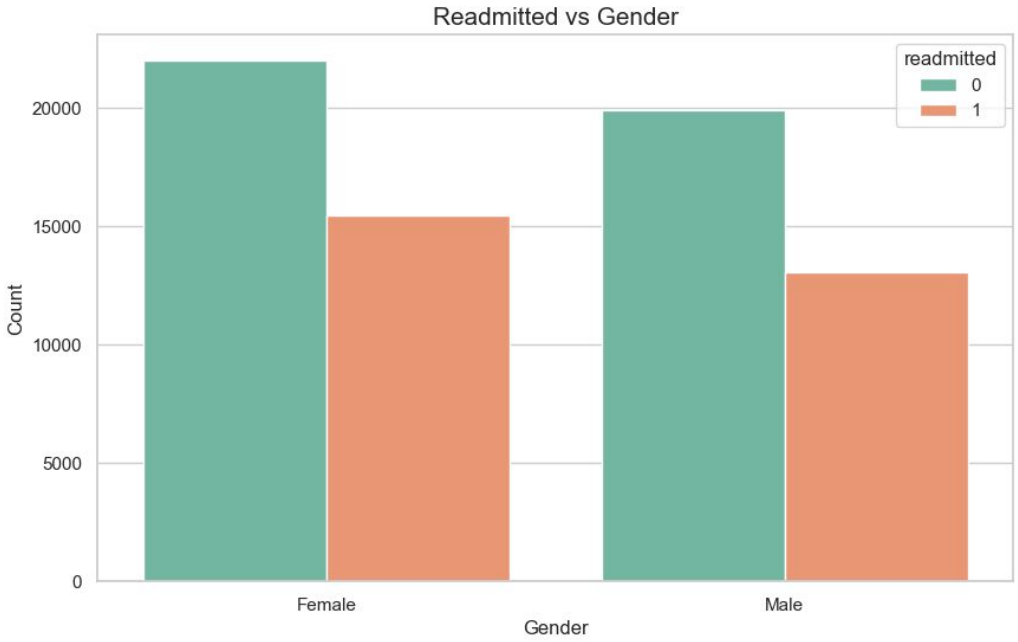
Visualizations

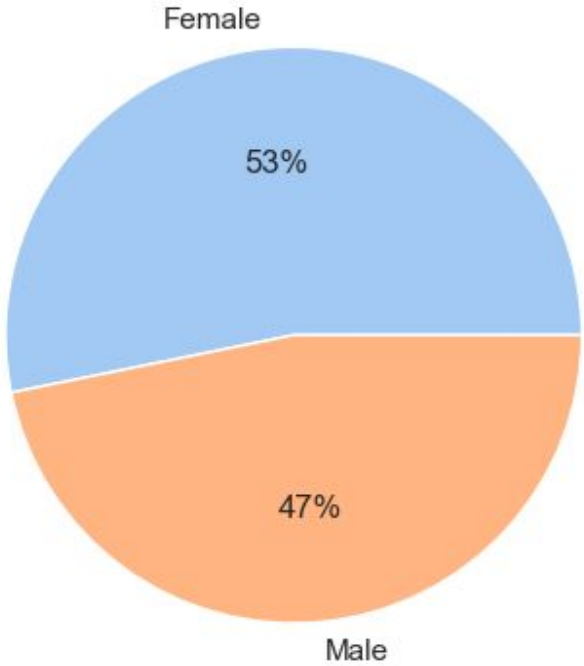




Patient readmission distribution

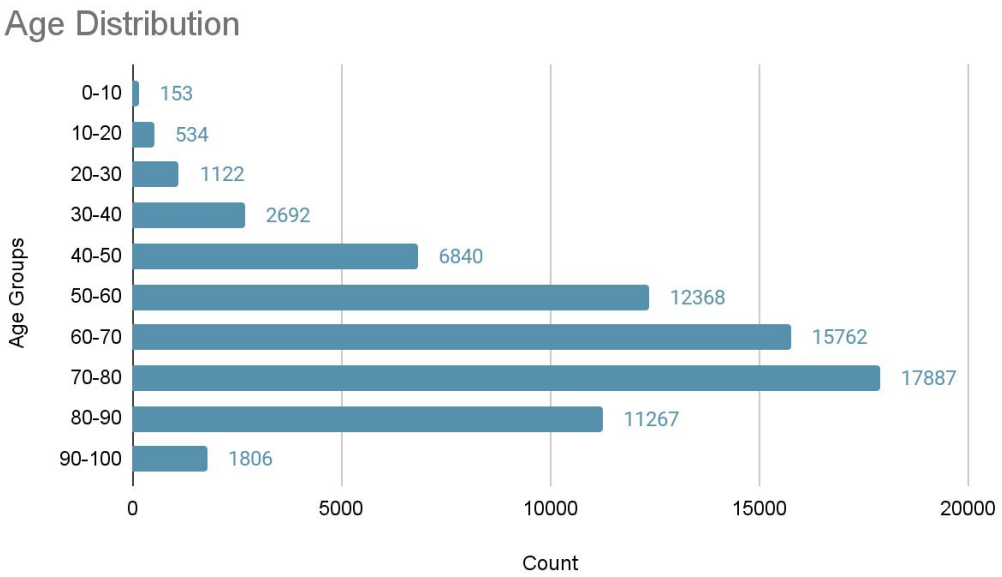
Visualizations





Gender wise distribution

Visualizations

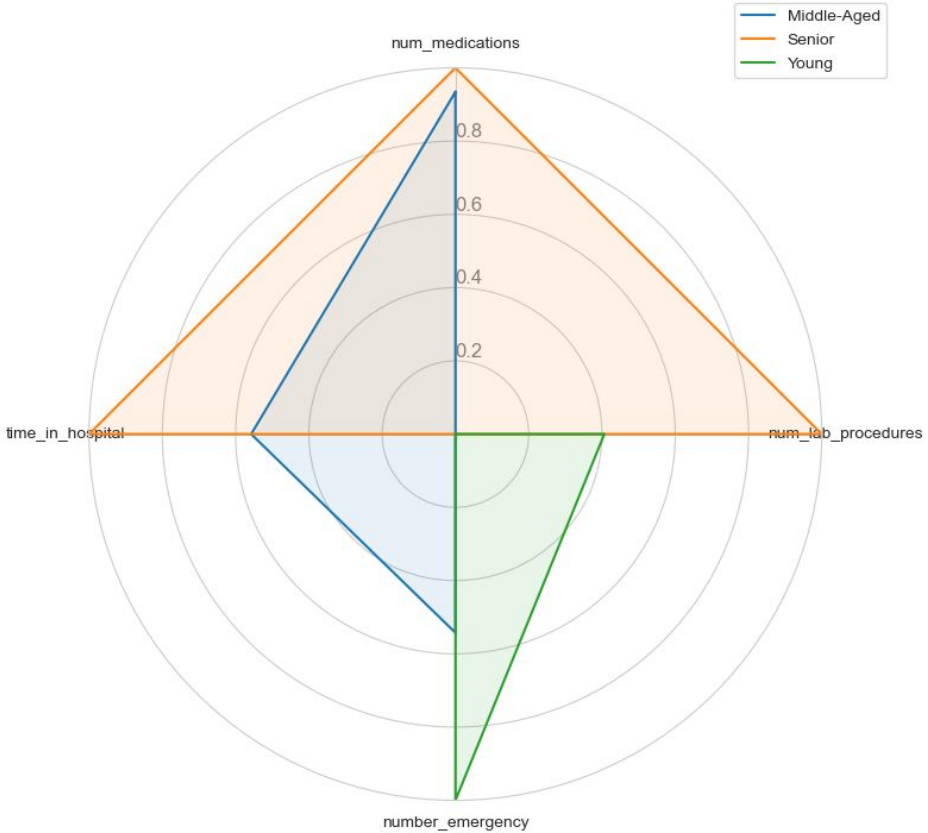


1. Senior Majority: The data shows a higher population count in the senior age brackets, especially for ages 70-80.

2. Fewer Youths: There is a lower count of younger age groups, with those under 20 being the least represented.

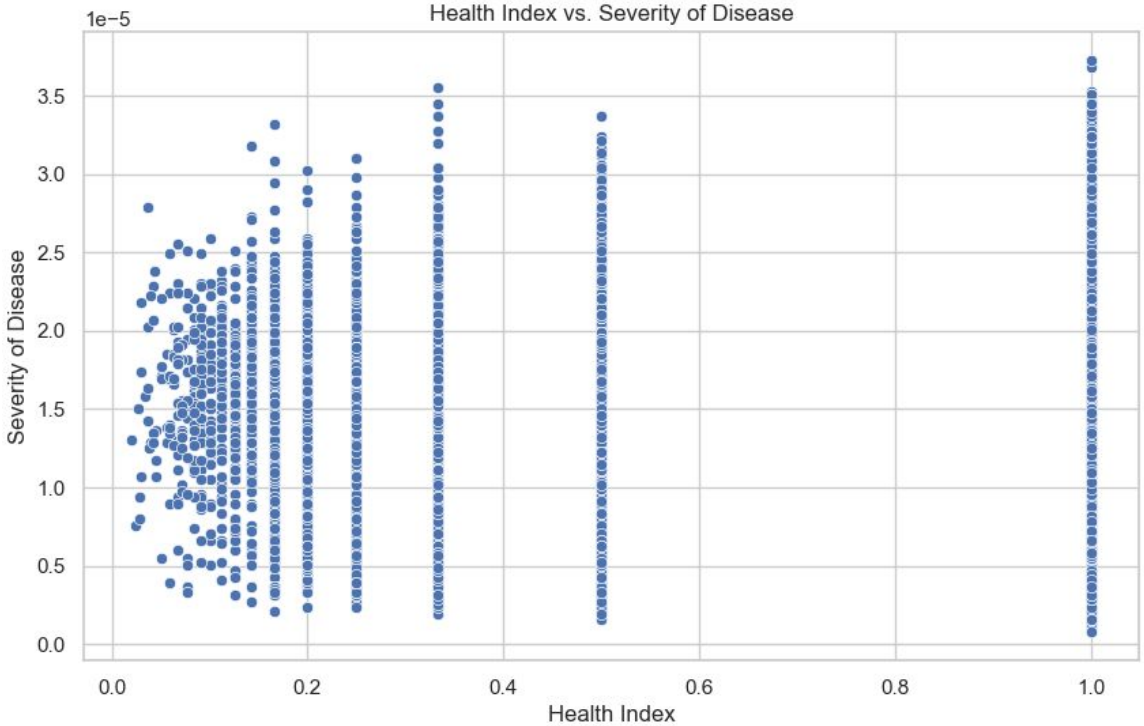
3. Moderate Middle Age Presence: Middle-aged adults (30-50 years) have a moderate presence, indicating a smaller but still notable portion of the workforce demographic.

Analysis of Healthcare Parameters Among Different Age Groups



- From the graph, it appears that Middle-Aged individuals have higher values for time in hospital and number of lab procedures compared to other groups.
- Seniors have the highest value for number of medications.
- Young individuals have notably higher values for number of emergencies.

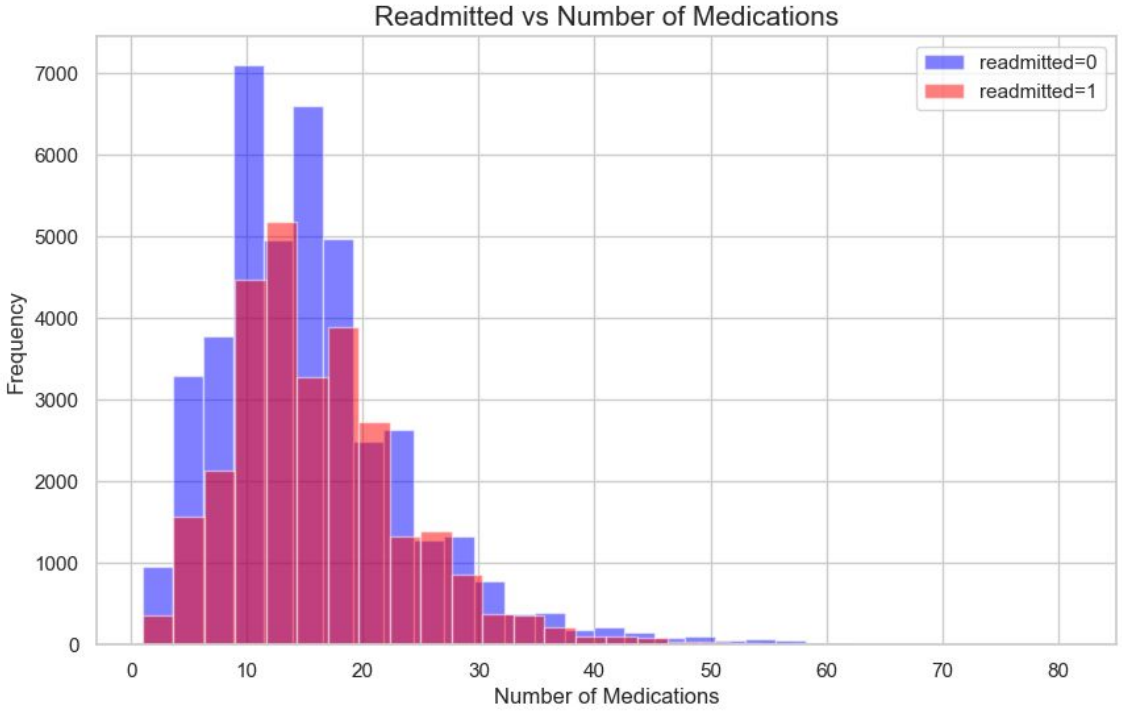
Analysis of Disease Severity vs Health Index



1. Data Clustering: Data points are heavily clustered at lower Health Index values, suggesting that within this population, a lower Health Index is more common.

2. Discrete Values: The Health Index data seems to be discrete or categorized into specific values rather than continuous, as indicated by the vertical lines of data points at regular intervals on the Health Index axis.

Analysis of Number of Medications

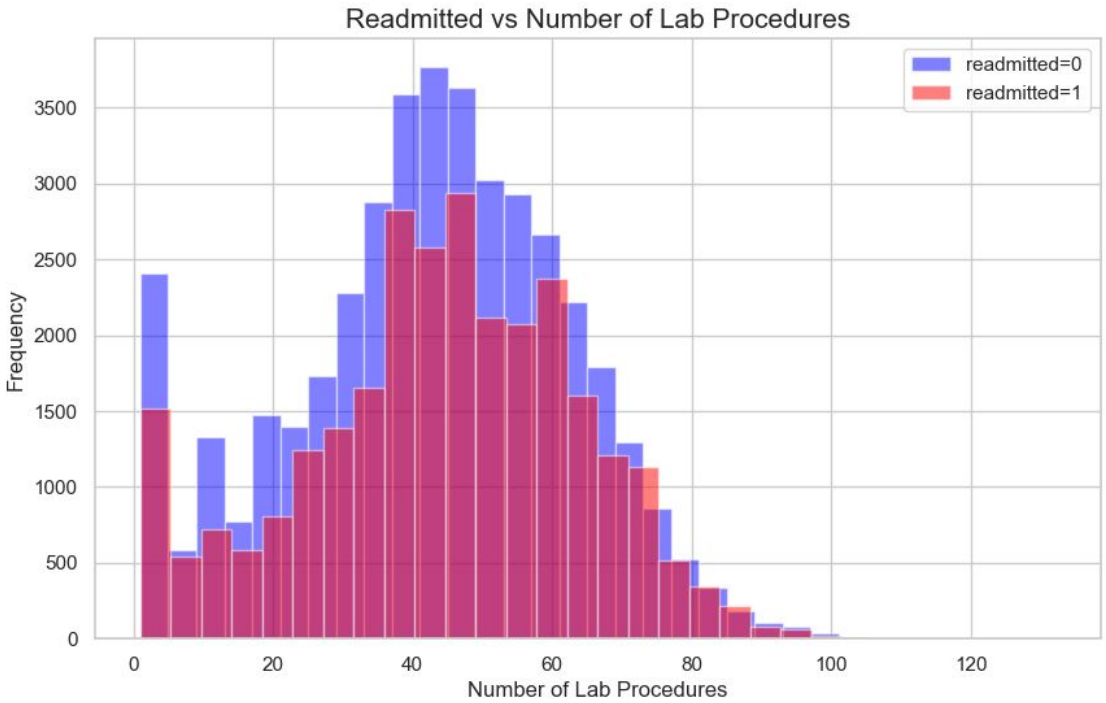


1. Common Prescription Range: Both readmitted and non-readmitted patients most commonly take 10-20 medications.

2. Lower Medication Counts: Non-readmitted patients predominantly fall within the lower medication count ranges.

3. Association with Readmission: A larger number of medications is more frequent among readmitted patients, suggesting a possible link to readmission rates.

Analysis of Number of Lab Procedures



1. Procedure Peaks: The data shows two peaks in lab procedures, indicating typical procedure ranges for patients.

2. Readmission Trends: Non-readmitted patients commonly have fewer lab procedures, while readmitted patients' procedures are more varied.

Patient Readmission Prediction

Data Preprocessing

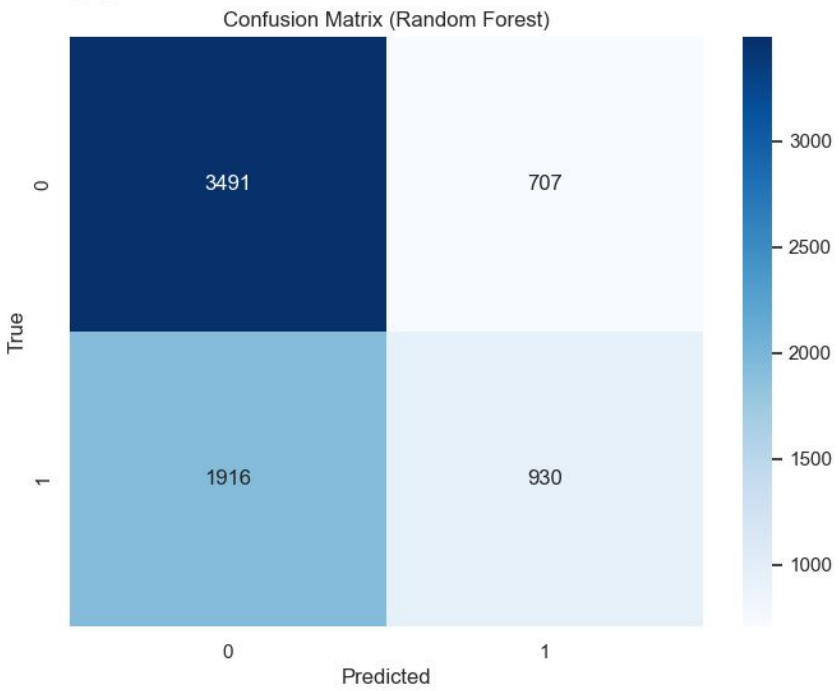
- An essential preprocessing step we undertook was the mapping of the ‘readmitted’ target column. Initially, this column had three distinct values: ‘<30’ for readmissions within 30 days, ‘>30’ for readmissions after 30 days, and ‘No’ for no readmissions.

- We processed these values in two distinct ways:
  - First: We mapped both ‘<30’ and ‘>30’ to 1, indicating readmission, and ‘No’ to 0, indicating no readmission.
  - Second: We mapped only ‘<30’ to 1, indicating readmission, and both ‘>30’ and ‘No’ to 0, indicating no readmission.

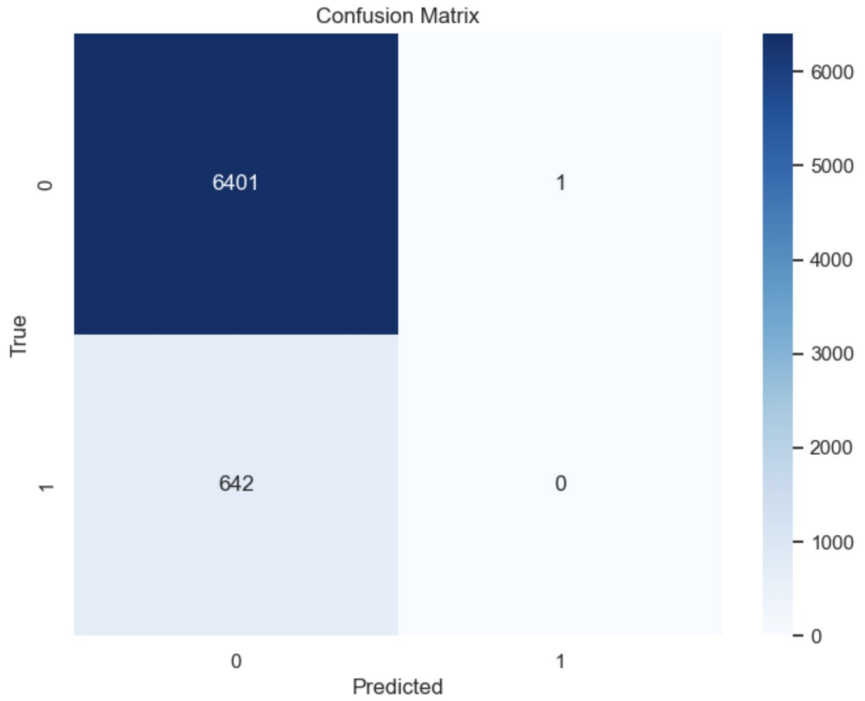
- These different mapping strategies significantly influenced the accuracy of our trained models, which we will discuss in the following slides.

Random Forest

First Method

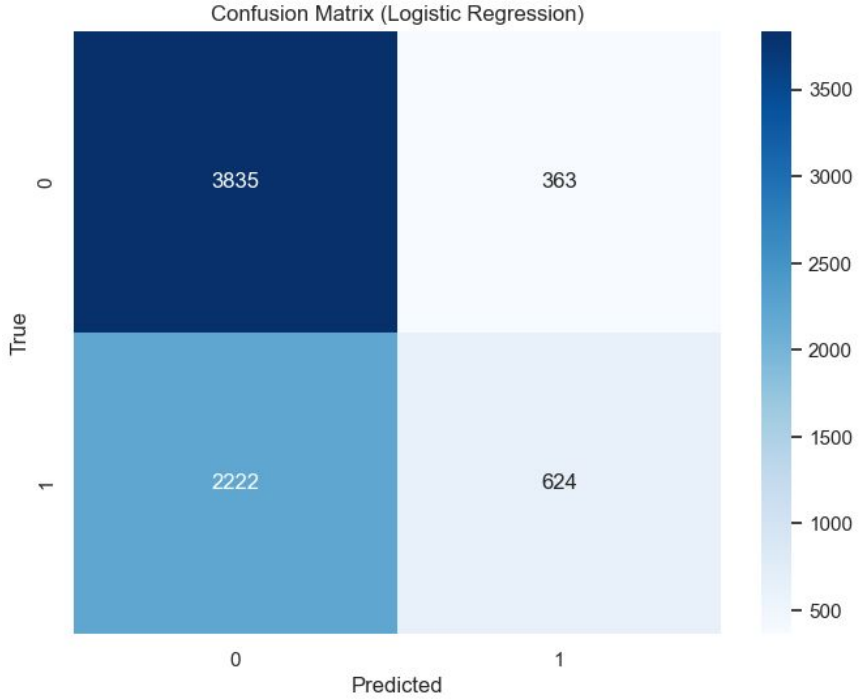


Second Method

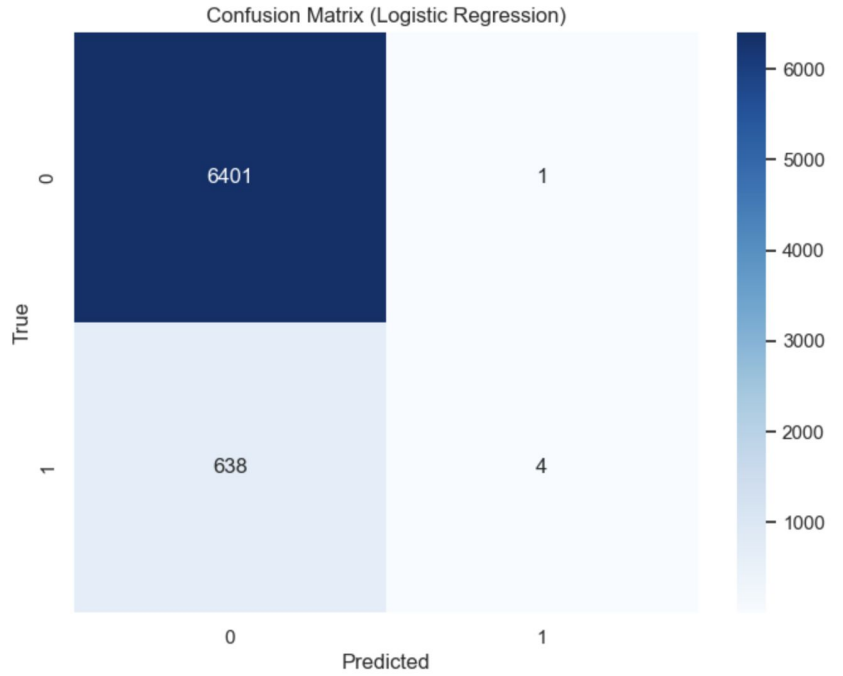


Logistic Regression

First Method

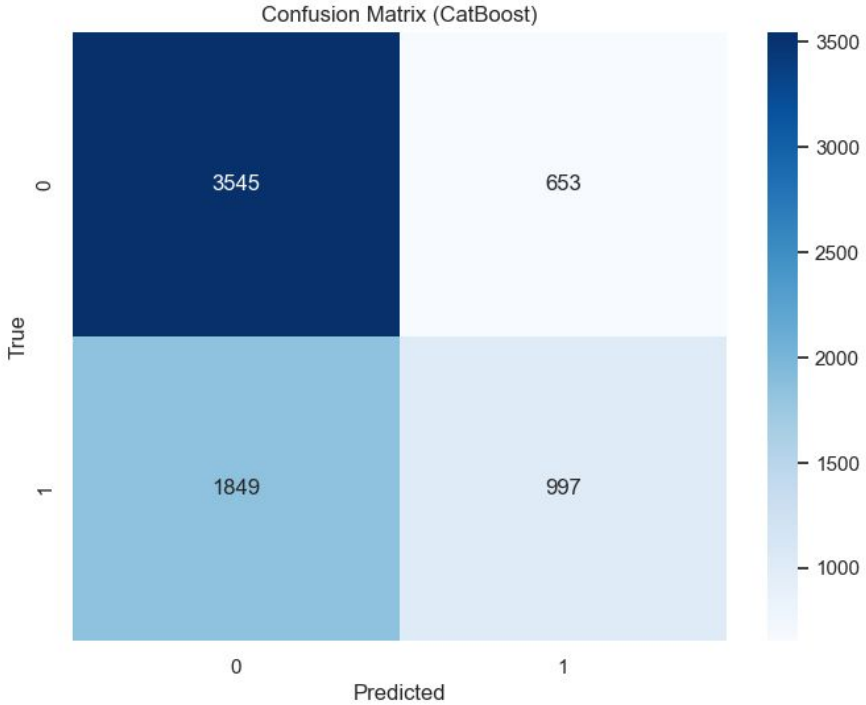


Second Method

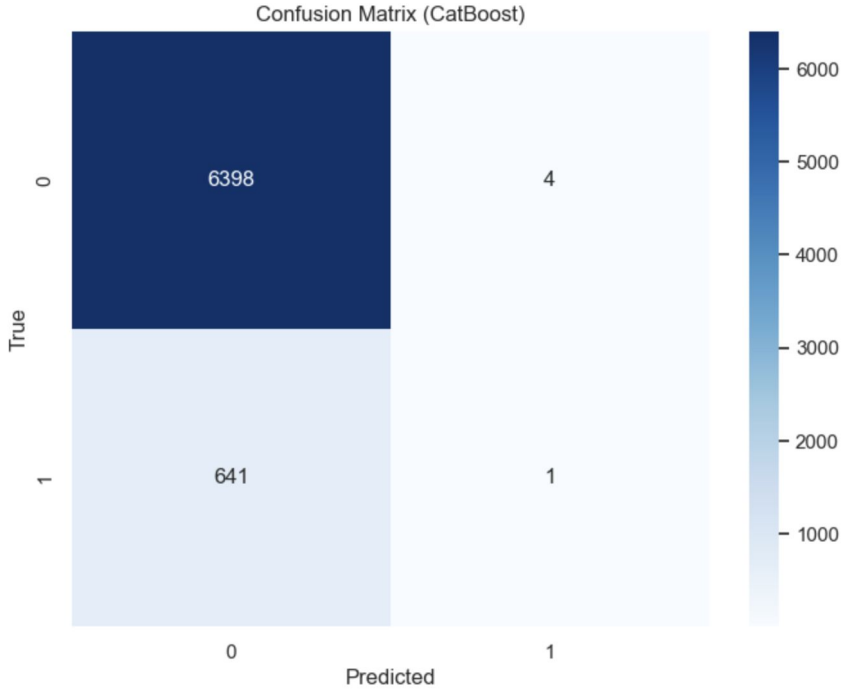


CatBoost

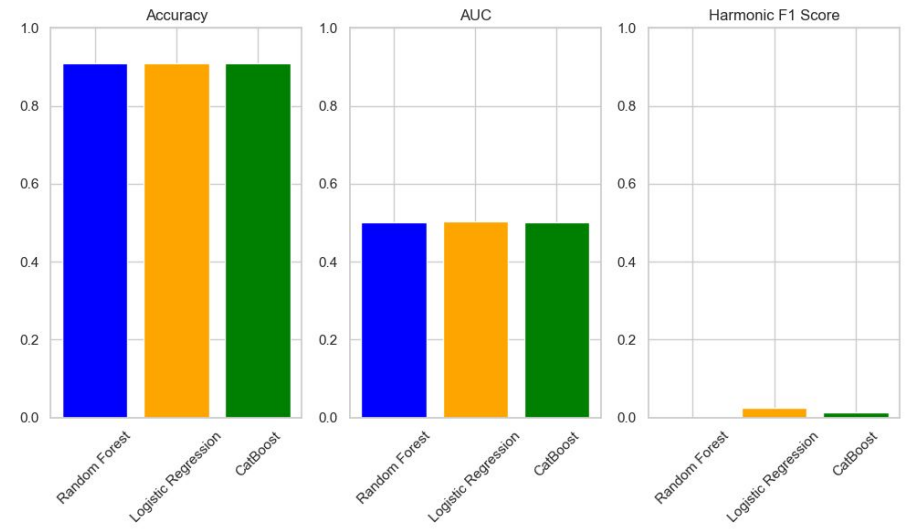
First Method

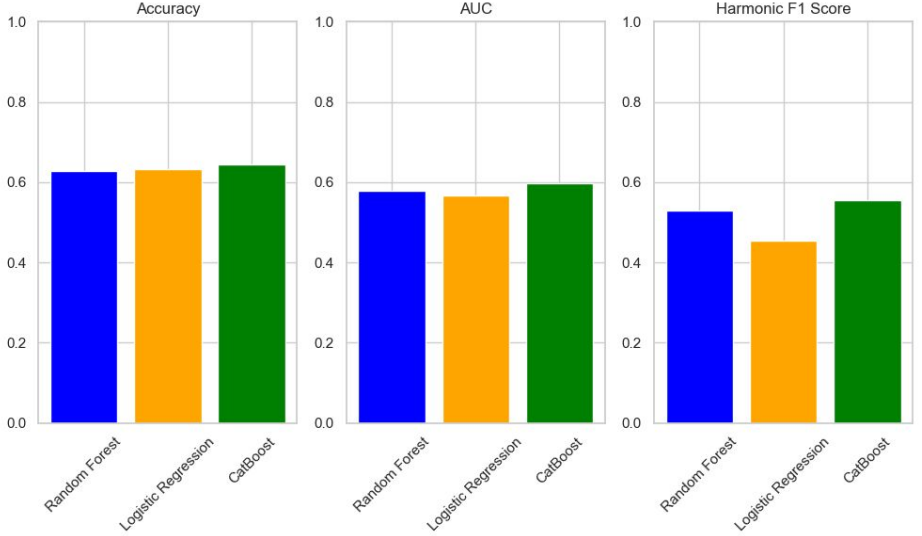


Second Method

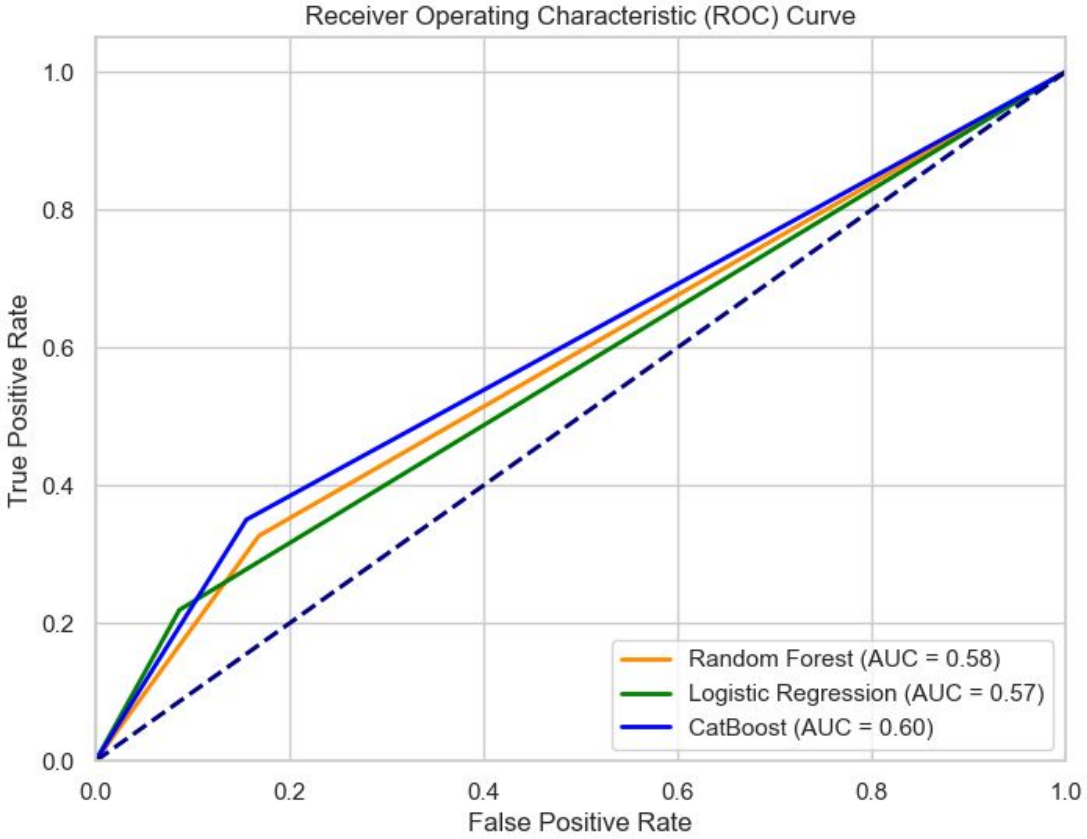


Model Metrics comparison





Model Metrics comparison



1. Best Model: CatBoost marginally outperforms others with the highest AUC of 0.60, indicating better classification capability.

2. Close Performance: All models show similar AUC scores, suggesting comparable effectiveness for the given classification task.

3. Complexity vs Performance: Despite CatBoost's superior AUC, its complexity requires considering factors like interpretability and computational cost.

Thank You