**AI Project Final Report**

**Hospital Readmission Prediction for Diabetic Patients**

**Submitted by:**

| **Student Name** | **Student ID** |
| --- | --- |
| محمد أحمد محمد عبد الغني | 2023170477 |
| نهى يوسف محمد موسى | 2023170669 |
| ملك رشاد محمود محمد | 2023170624 |
| ارسانيوس نبيل نيري عطية | 2023170081 |
| إيريني عادل إسحق سعيد | 2023170114 |
| سمية محمود حسنين ابراهيم | 2023170276 |
| يحيى أحمد يحيى فايز | 2023170704 |

**Supervised by:**

**Dr. Mohamed Magdy**

**Project Overview :**

This project aims to build a predictive system that identifies whether a diabetic patient is likely to be **readmitted to the hospital within 30 days** after discharge. The data used comes from a large, real-world medical dataset containing **101,766 patient records**, each with **50 features** related to demographics, lab results, diagnoses, medications, and prior hospital visits.

This is a **Supervised Machine Learning project** because:

* The data is **labeled** — each patient record includes a readmitted value that tells us the outcome (0 if the patient was readmitted before 30 days, 1 if he was readmitted after 30 days, 2 if not).
* Our goal is to **learn a pattern from these labeled examples** and use it to predict the label (readmission status) for new patients.
* Prediction
  + Predict one of three labels:
  + "<30" : 0
  + ">30" : 1
  + "NO" : 2

**1.Data Preprocessing:**

* 1. **Initial Inspection**
* Decoded id columns using mapping file. (ID’s mapping file).
* Detected both invalid and null values in several features.
* For (diag1,diag2,diag3) we first decoded them according to “ICD-9” <https://en.wikipedia.org/wiki/List_of_ICD-9_codes>.
  1. **Handling Missing & Invalid Values**
* Invalid (non-null but wrong) values were also identified and replaced with the null.
* Null values were replaced using mode imputation (most frequent value).

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

* All missing values were filled by mode except (race and payer code) were dropped.
* The weight feature contained mostly nulls, so a predictive Random Forest model was trained using age and gender to estimate the missing values.
  1. **Encoding Categorical Features**
* All categorical features were encoded using:
  + One-Hot Encoding (for nominal features).

(Age, weight,diag1,diag2,diag3).

* + Label Encoding (for ordinal features or when applicable).

(Gender, admission\_type\_id, discharge\_disposition\_id

admission\_source\_id, medical\_specialty, insulin

readmitted, metformin, max\_glu\_serum,

A1Cresult, repaglinide, nateglinide,

glimepiride, glipizide, glyburide,

pioglitazone, rosiglitazone, glyburide-metformin

change, diabetesMed).

* All mappings from this process were saved in a structured JSON file for reproducibility and future decoding.

**1.4 Dropping Irrelevant Features**

* Some features were dropped based on their distribution of value:
  + Extremely low variance.

A black screen with white text

AI-generated content may be incorrect.A black screen with white text

AI-generated content may be incorrect.

A computer screen shot of a black screen

AI-generated content may be incorrect.A screen shot of a computer program

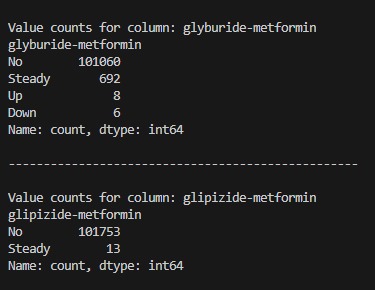
AI-generated content may be incorrect.A black background with white text

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.A computer screen with white text

AI-generated content may be incorrect.A computer screen shot of a black screen

AI-generated content may be incorrect.

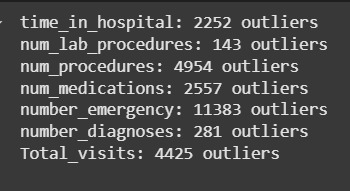


* + Too many unique values.

A screenshot of a computer

AI-generated content may be incorrect.

* + We combined “number\_outpatient” and “number\_inpatient” and “number\_emergency” into “Total\_visits.”
  + We combined all the weights above 125 into one column “weight\_>\_125”.
  1. **Outliers Handling**
* Capping technique was applied to numerical features to limit the impact of extreme values.
* Number of outliers in the numerical features:



* 1. **Standardization**
* Standard normalization (z-score normalization) was applied to scale the data. This method is broadly applicable and helps ensure that all features contribute equally during model training, regardless of their original scales.
* The transformation is defined as:

**2. Exploratory Data Analysis (EDA):**

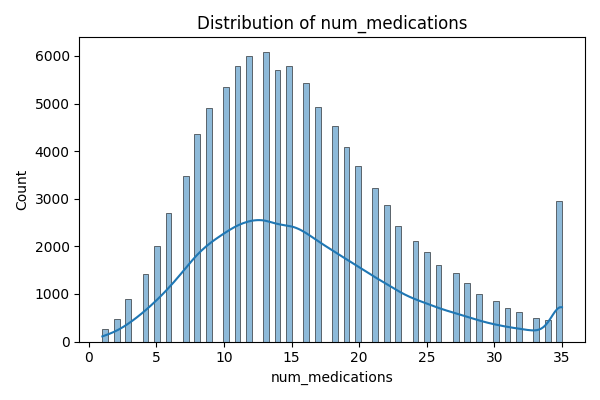
To better understand the dataset, we categorized the features into five key groups:

1. Numerical Features
2. Categorical Features
3. Medical / Treatment Features
4. Diagnosis Features
5. Feature Correlation & Impact

This modular structure allowed us to **investigate patterns**, trends, and their relationship with hospital readmission in a focused and interpretable way.

**2.1 Numerical Features:**

**Includes**: time\_in\_hospital , num\_lab\_procedures, num\_procedures, Total\_visits ,etc.



**2.2Categorical Features :**

* Analyzed distributions using **count** plots and **grouped** bar charts.
* Converted most of these to numeric using label or one-hot encoding.
* Categorical variables like admission\_type\_id and discharge\_disposition\_id showed some **patterns** with readmitted.

A graph of a number of patients

AI-generated content may be incorrect.

**2.3 Medication Features:**

* These columns indicate how diabetes-related medications were used (up, down, steady, no).
* Medication change status was encoded and analyzed across readmitted vs. non-readmitted patients.

**A graph of different colored bars

AI-generated content may be incorrect.**

**2.4 Diagnosis Features:**

We grouped diagnosis data into **three categories** based on the original diagnosis columns :(d1,d2 and d3) :-

* These were binary features indicating the presence of each diagnosis category.
* We analyzed their frequencies in readmitted vs. not-readmitted patients.

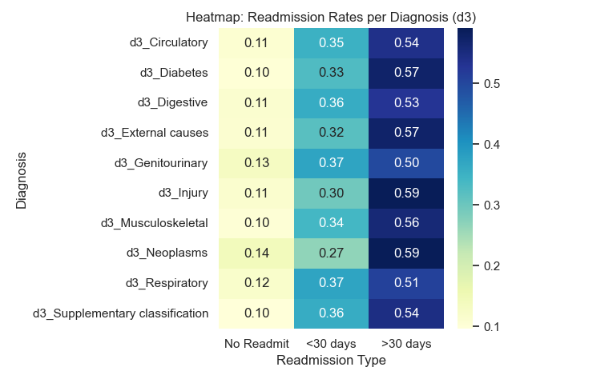
A graph of a number of diseases

AI-generated content may be incorrect.

**2.5 Feature Correlation & Importance**

To understand feature relationships and identify the most impactful ones, we:

* Computed **correlation matrices** for:
  + Numerical features
  + Medication usage
  + Diagnosis category flags
* Used **tree-based models (e.g., Random Forest)** to extract **feature importance**.



A screenshot of a graph

AI-generated content may be incorrect.

A chart of a patient's health

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

A diagram of a heat map

AI-generated content may be incorrect.

**Summary:**

By dividing the features into **Num / Cat / Med / Diag (1,2,3)** we performed a comprehensive and structured exploration. This allowed our team to isolate meaningful patterns and relationships critically for building a strong predictive model.

**3. Model Development :**

Before starting training, we prepared the data by:

1. **Feature Selection:**

* Use SelectKBest(f\_classif) to select the top 15 features.

1. **Handling Class Imbalance:**

* Used RandomOverSampler to balance class distribution before training.

1. **Dataset Splitting**

* Split into training (80%) and testing (20%) sets using stratified sampling to preserve class proportions.

1. **Model Training:**

* **Train the following models:**
  + **Logistic Regression**
  + **SVM (Linear Kernel)**
  + **Random Forest (150 Trees)**
  + **XGBoost** with class weights ({0:3, 1:2, 2:2})

1. **Model Evaluation:**

* Predict and evaluate each model.
* Metrics used:

Classification Report (Precision, Recall, F1-Score):

A screenshot of a computer screen

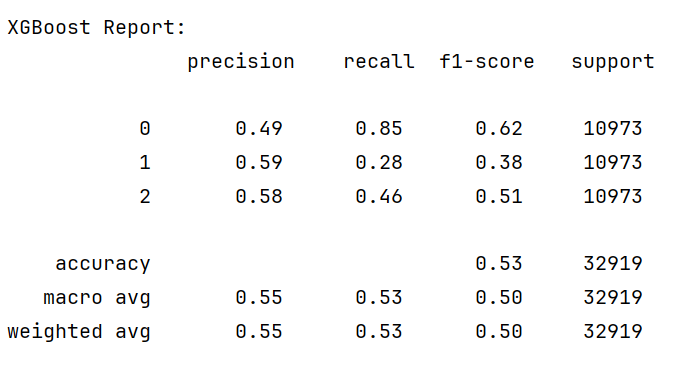
AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.



* ROC-AUC Score (macro-averaged, multi-class):

A close up of black text

AI-generated content may be incorrect.

**6.Confusion Matrix Visualization:**

* Plots for all models to visualize performance across classes.

A graph of confusion matrix

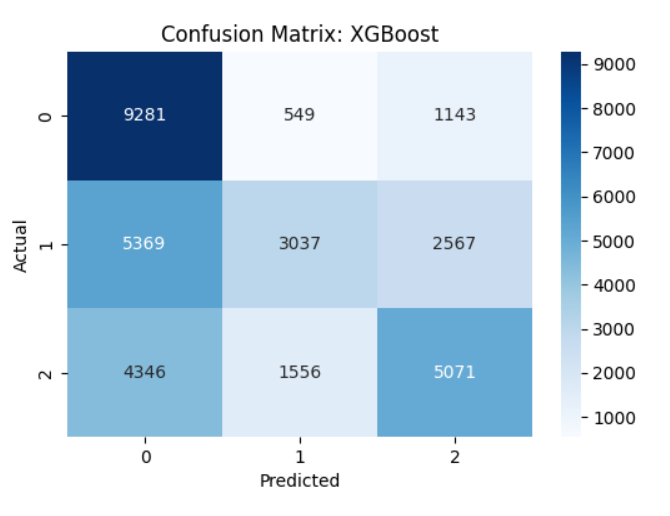
AI-generated content may be incorrect.

A blue and white graph

AI-generated content may be incorrect.

A graph with numbers and a number in blue squares

AI-generated content may be incorrect.



**Summary:**

* Random Forest significantly outperformed all other models, both in accuracy and ROC-AUC, making it the most reliable for readmission prediction.
* Its ensemble nature allowed it to capture complex interactions between patient features.
* Models like Logistic Regression and SVM were limited by their linear nature and lack of robustness to class imbalance.
* XGBoost performed reasonably well but required intensive tuning and still fell short of Random Forest's consistency.