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

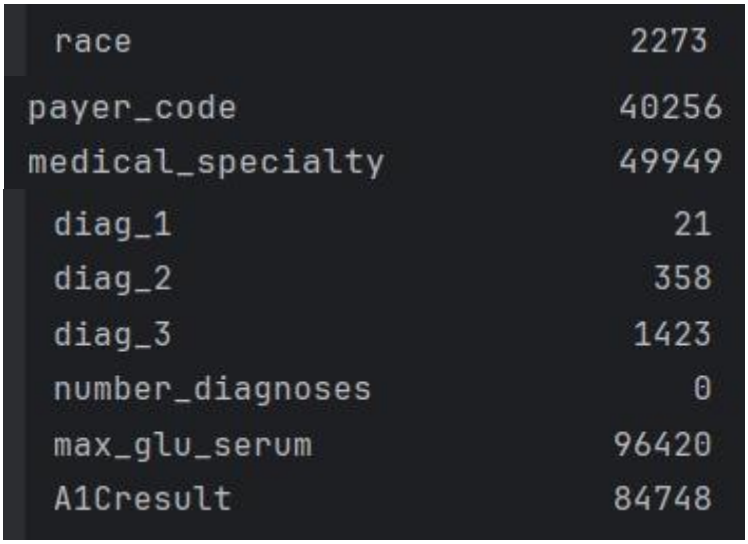
1.Data Preprocessing:

1.1 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.

1.2 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).



race	2273
payer_code	40256
medical_specialty	49949
diag_1	21
diag_2	358
diag_3	1423
number_diagnoses	0
max_glu_serum	96420
A1Cresult	84748

- 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.

1.3 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.

```
Value counts for column: acetohexamide
acetohexamide
No        101765
Steady         1
Name: count, dtype: int64
```

```
Value counts for column: citoglipton
citoglipton
No    101766
Name: count, dtype: int64
```

```
Value counts for column: miglitol
miglitol
No        101728
Steady        31
Down           5
Up             2
Name: count, dtype: int64
```

```
metformin-pioglitazone
No        101765
Steady         1
Name: count, dtype: int64
```

```
glipizide-metformin
No        101753
Steady        13
Name: count, dtype: int64
```

```
Value counts for column: troglitazone
troglitazone
Steady         3
Name: count, dtype: int64

----------------------------------------

Value counts for column: tolazamide
tolazamide
No        101727
Steady        38
Up             1
Name: count, dtype: int64

----------------------------------------

Value counts for column: examide
examide
No    101766
Name: count, dtype: int64
```

```
Value counts for column: acarbose
acarbose
No        101458
```

```
glimepiride-pioglitazone
No        101765
Steady         1
Name: count, dtype: int64
```

```
Value counts for column: metformin-rosiglitazone
metformin-rosiglitazone
No        101764
Steady         2
Name: count, dtype: int64
```

- Too many unique values.

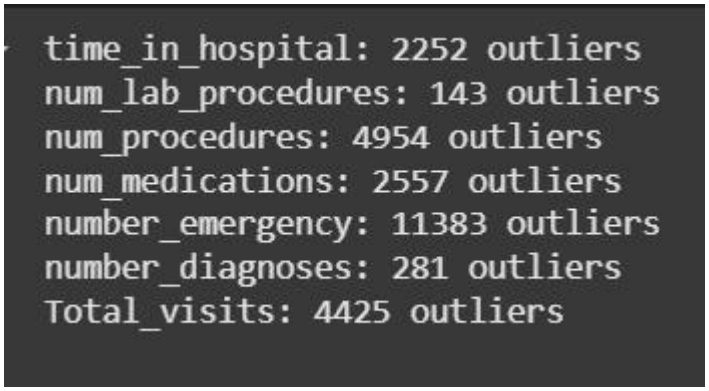
```
Value counts for column: patient_nbr
patient_nbr
88785891     40
43140906     28
88227540     23
1660293      23
23199021     23
             ..
174477542     1
38726739      1
77391171      1
89869032      1
63555939      1
Name: count, Length: 71518, dtype: int64
```

- 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”.

1.4 Outliers Handling

- Capping technique was applied to numerical features to limit the impact of extreme values.
- Number of outliers in the numerical features:



1.5 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:

$$z = \frac{\{x - \mu\}}{\{\sigma\}}$$

2. Exploratory Data Analysis (EDA):

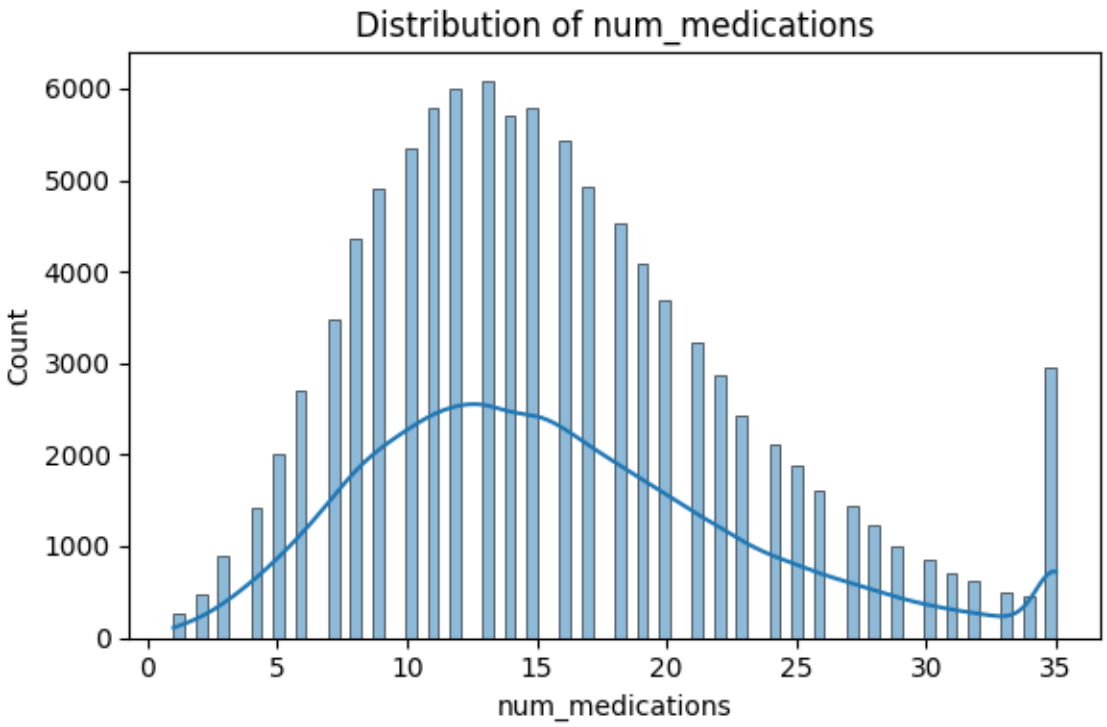
To better understand the dataset, we categorized the features into five key groups:

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.

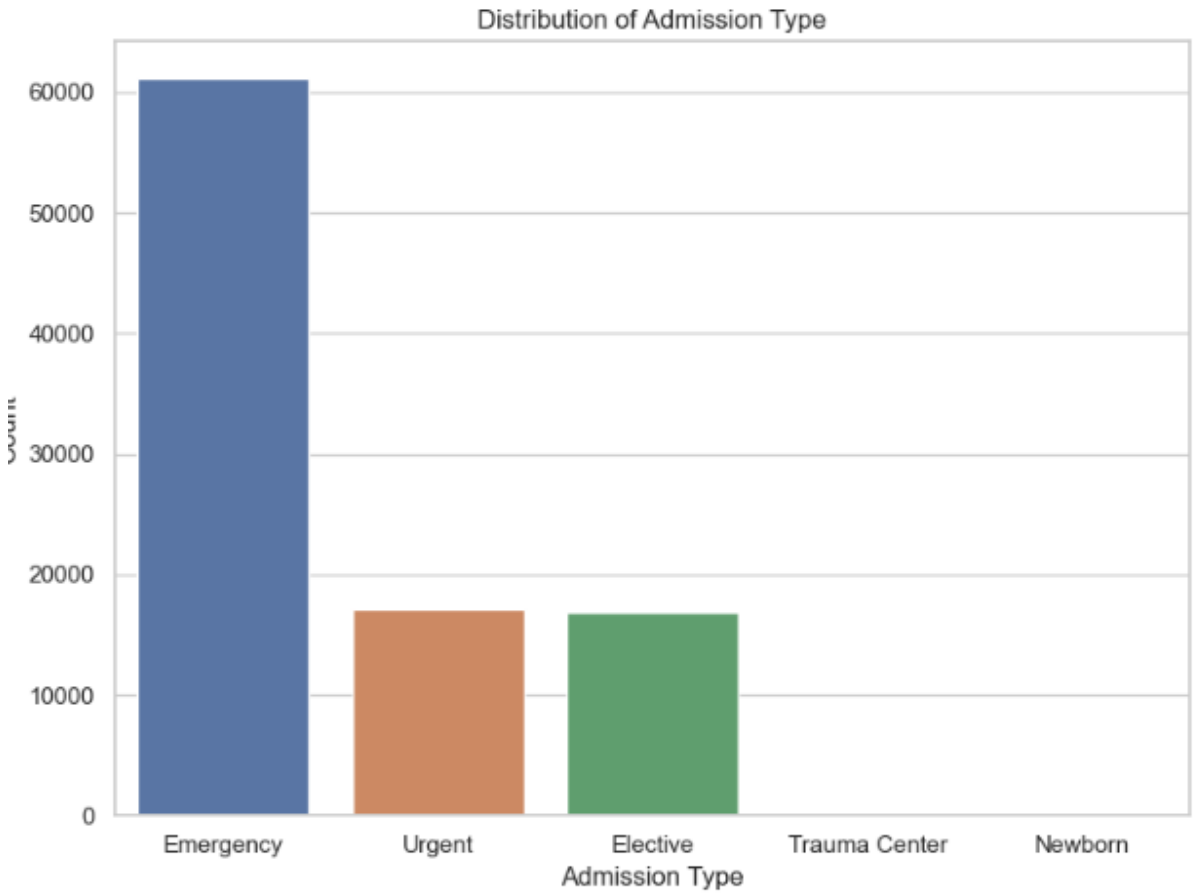
2.1 Numerical Features:

Includes: time_in_hospital , num_lab_procedures, num_procedures, Total_visits ,etc.



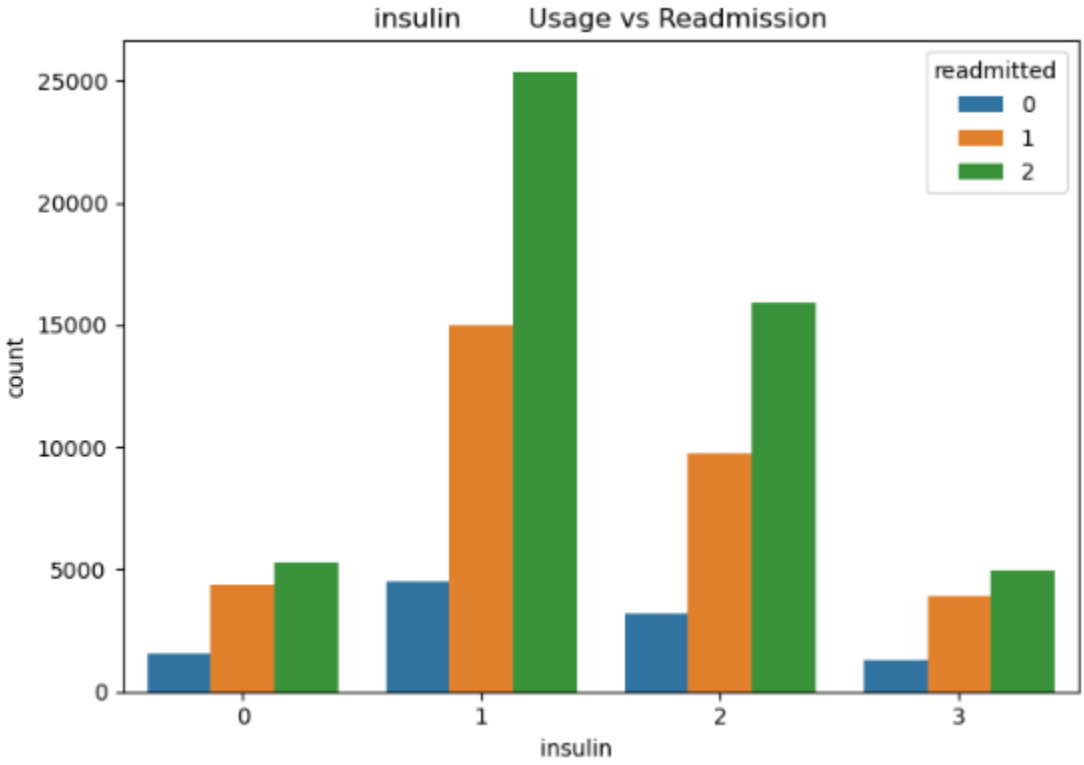
2.2Categorical Features :

- 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.
-



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.

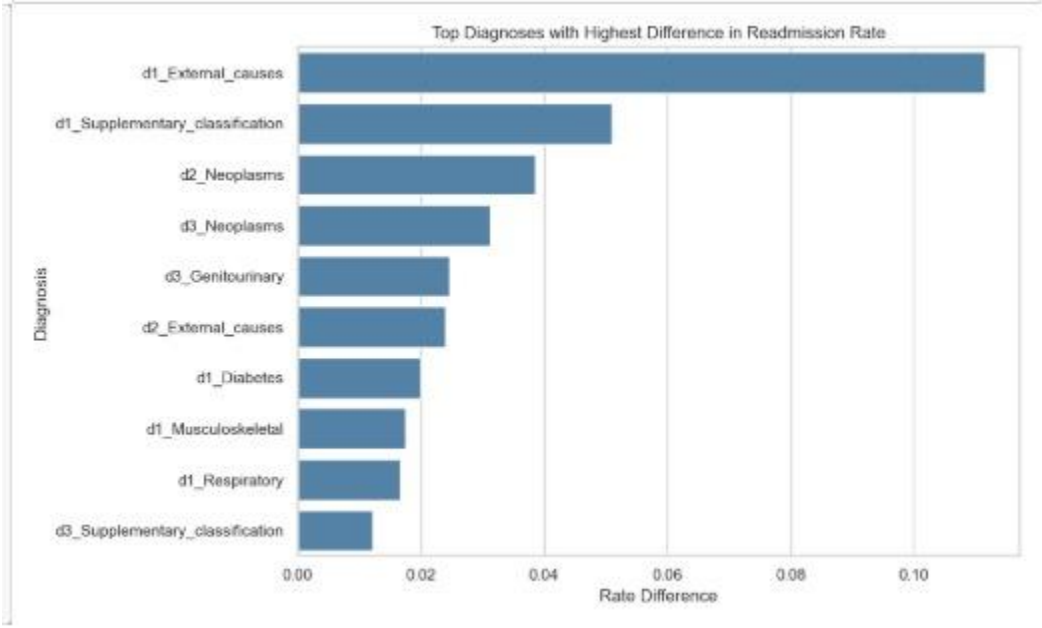


2.4 Diagnosis Features:

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.

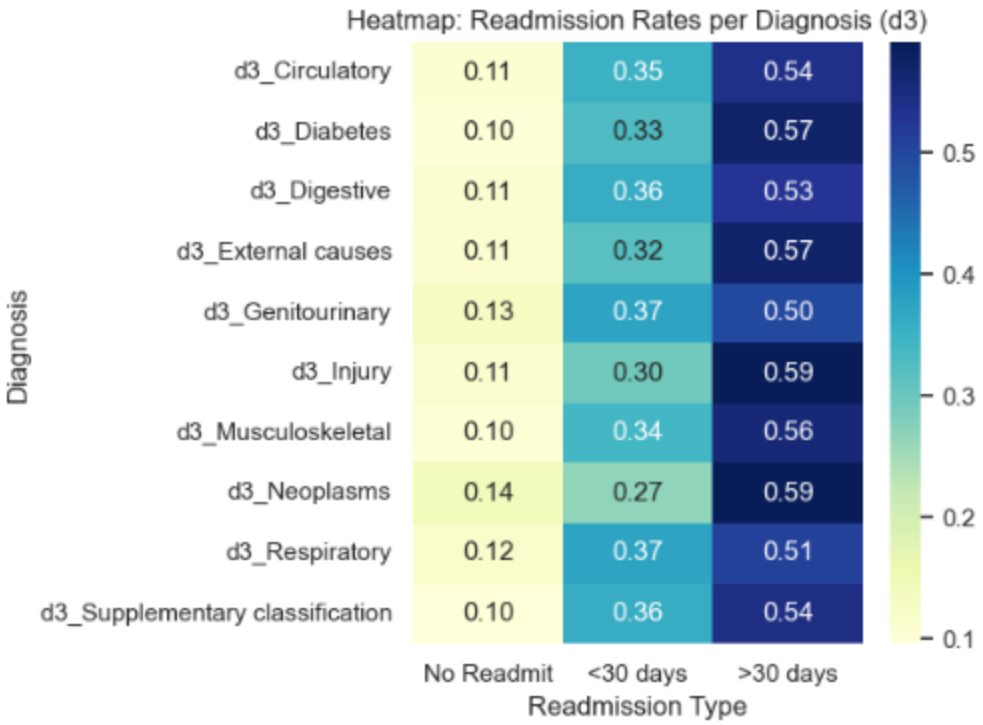
- We analyzed their frequencies in readmitted vs. not-readmitted patients.

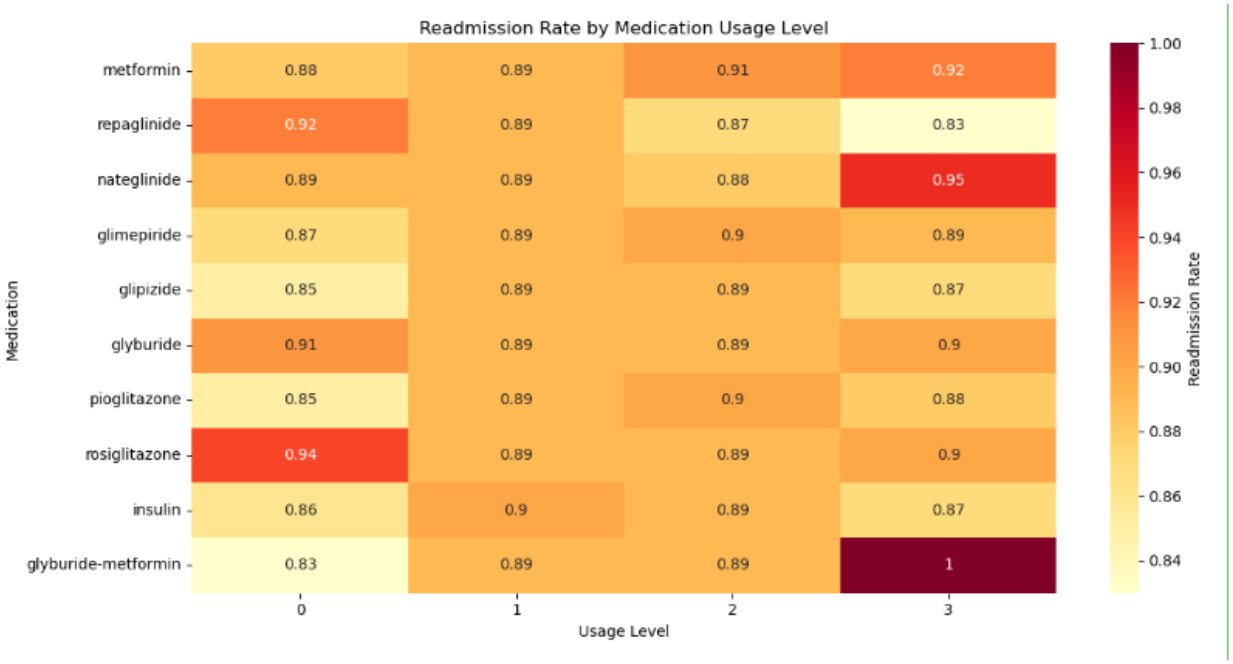


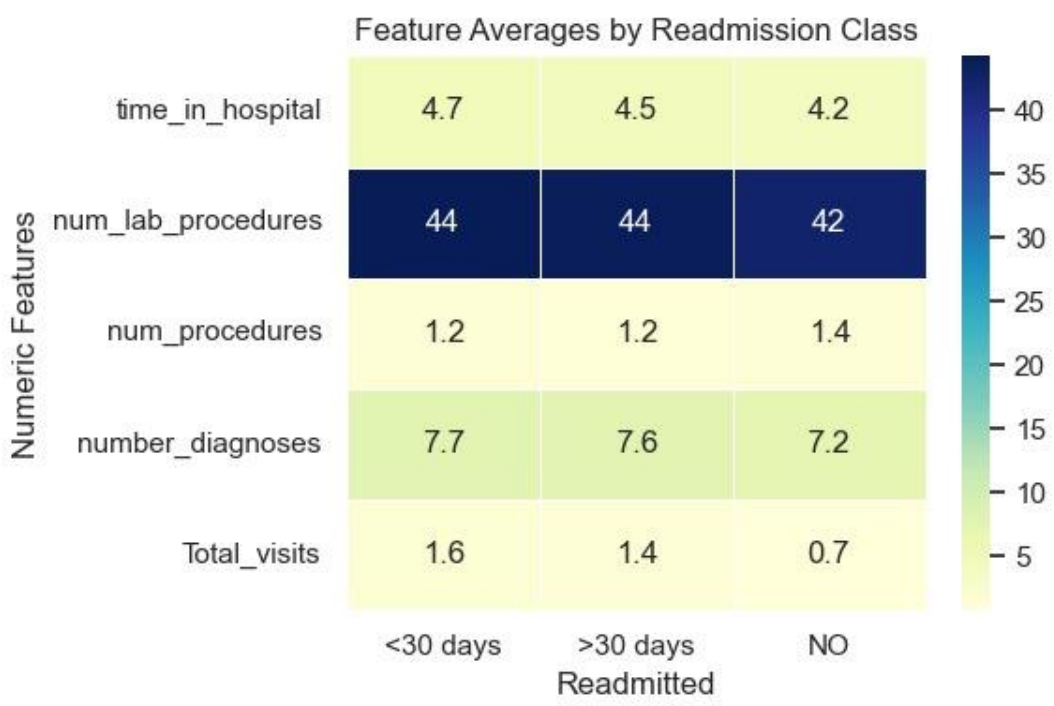
2.5 Feature Correlation & Importance

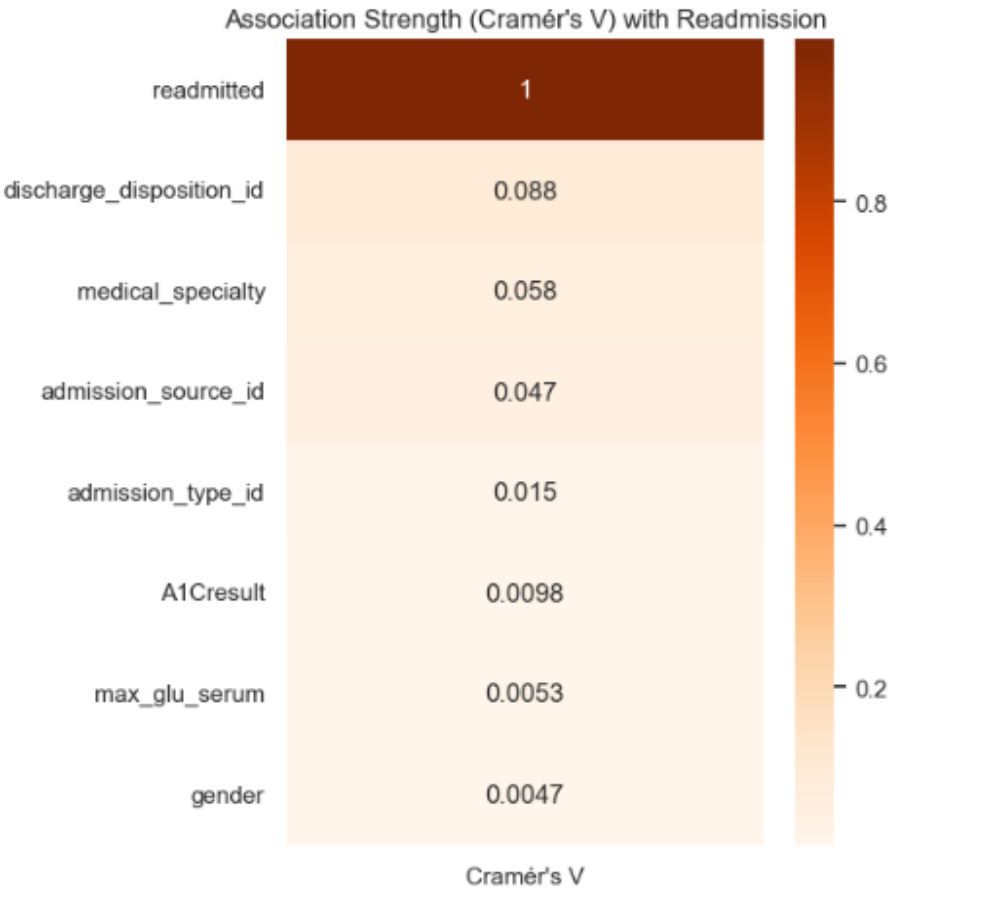
To understand feature relationships and identify the most impactful ones, we:

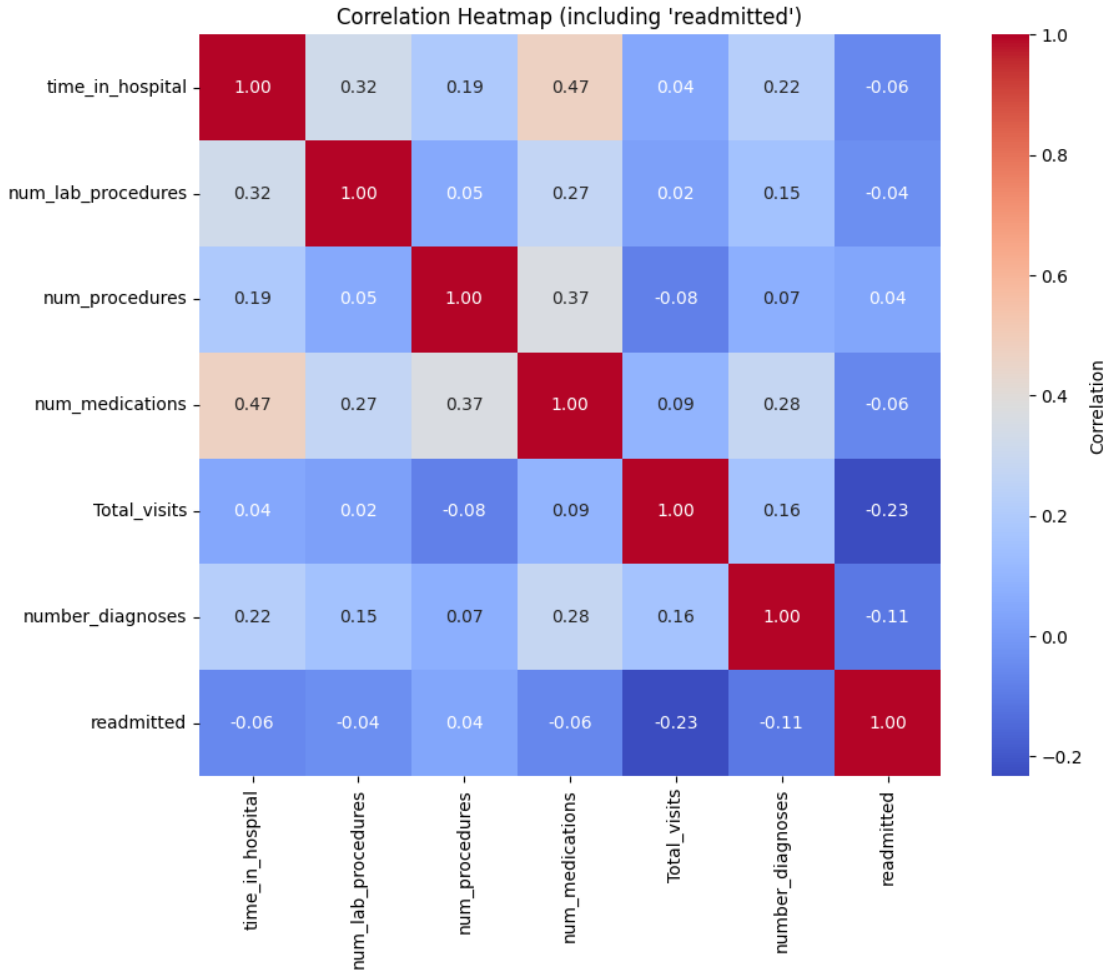
- Computed **correlation matrices** for:
  - Numerical features
  - Medication usage
  - Diagnosis category flags
- Used **tree-based models (e.g., Random Forest)** to extract **feature importance**.


Heatmap: Readmission Rates per Diagnosis (d3)
Diagnosis
d3_Circulatory 0.11 0.35 0.54
d3_Diabetes 0.10 0.33 0.57
d3_Digestive 0.11 0.36 0.53
d3_External causes 0.11 0.32 0.57
d3_Genitourinary 0.13 0.37 0.50
d3_Injury 0.11 0.30 0.59
d3_Musculoskeletal 0.10 0.34 0.56
d3_Neoplasms 0.14 0.27 0.59
d3_Respiratory 0.12 0.37 0.51
d3_Supplementary classification 0.10 0.36 0.54
0.5
0.4
0.3
0.2
0.1
No Readmit
<30 days
>30 days
Readmission Type


Readmission Rate by Medication Usage Level
Medication
metformin 0.88 0.89 0.91 0.92
repaglinide 0.92 0.89 0.87 0.83
nateglinide 0.89 0.89 0.88 0.95
glimepiride 0.87 0.89 0.9 0.89
glipizide 0.85 0.89 0.89 0.87
glyburide 0.91 0.89 0.89 0.9
pioglitazone 0.85 0.89 0.9 0.88
rosiglitazone 0.94 0.89 0.89 0.9
insulin 0.86 0.9 0.89 0.87
glyburide-metformin 0.83 0.89 0.89 1
0
1
2
3
Usage Level
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
Readmission Rate


Feature Averages by Readmission Class
Numeric Features
time_in_hospital
num_lab_procedures
num_procedures
number_diagnoses
Total_visits
4.7
4.5
4.2
44
44
42
1.2
1.2
1.4
7.7
7.6
7.2
1.6
1.4
0.7
40
35
30
25
20
15
10
5
<30 days
>30 days
NO
Readmitted


Association Strength (Cramér's V) with Readmission
readmitted
discharge_disposition_id
medical_specialty
admission_source_id
admission_type_id
A1Cresult
max_glu_serum
gender
1
0.088
0.058
0.047
0.015
0.0098
0.0053
0.0047
0.8
0.6
0.4
0.2
Cramér's V



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.

2. **Handling Class Imbalance:**
   - Used RandomOverSampler to balance class distribution before training.

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

4. **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})

5. Model Evaluation:

- Predict and evaluate each model.
- Metrics used:

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

Logistic Regression Report:

	precision	recall	f1-score	support
0	0.43	0.41	0.42	10973
1	0.38	0.21	0.27	10973
2	0.44	0.65	0.52	10973
accuracy			0.42	32919
macro avg	0.42	0.42	0.41	32919
weighted avg	0.42	0.42	0.41	32919

SVM Report:

	precision	recall	f1-score	support
0	0.31	0.41	0.36	10973
1	0.33	0.12	0.18	10973
2	0.30	0.40	0.35	10973
accuracy			0.31	32919
macro avg	0.32	0.31	0.29	32919
weighted avg	0.32	0.31	0.29	32919

Random Forest Report:

	precision	recall	f1-score	support
0	0.94	0.99	0.97	10973
1	0.74	0.90	0.81	10973
2	0.87	0.64	0.74	10973
accuracy			0.84	32919
macro avg	0.85	0.84	0.84	32919
weighted avg	0.85	0.84	0.84	32919

XGBoost Report:

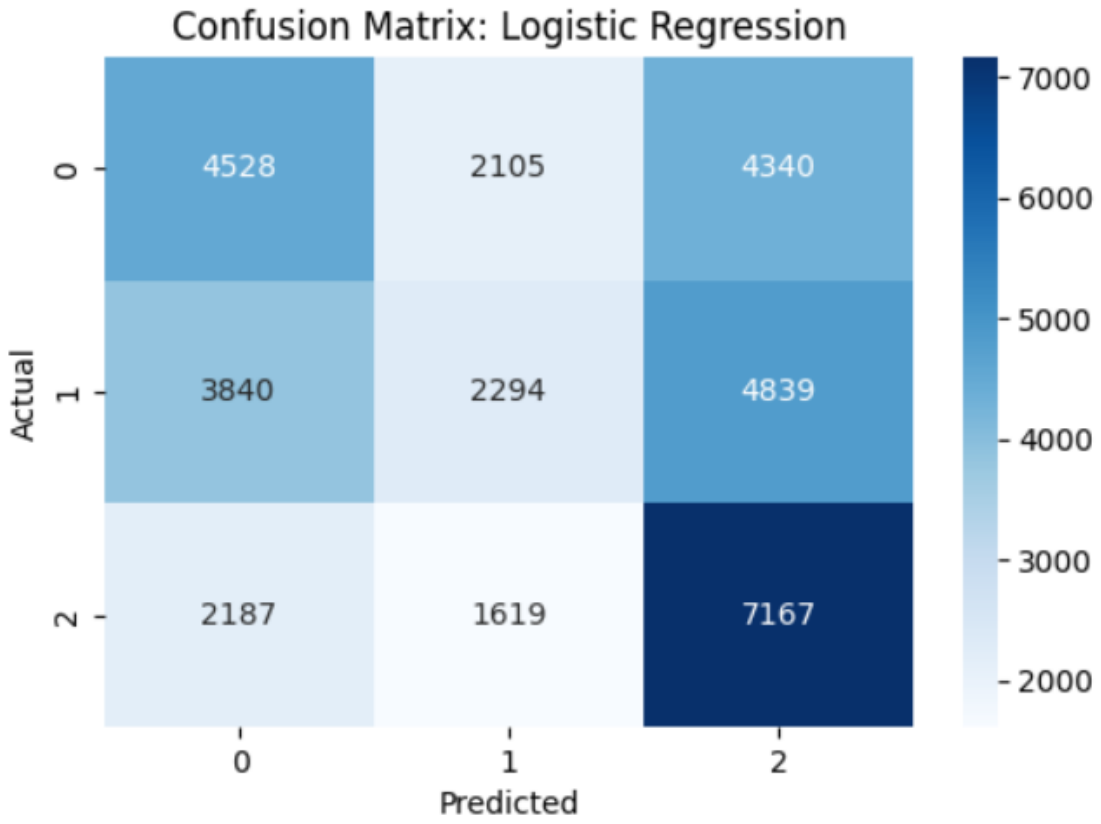
	precision	recall	f1-score	support
0	0.49	0.85	0.62	10973
1	0.59	0.28	0.38	10973
2	0.58	0.46	0.51	10973
accuracy			0.53	32919
macro avg	0.55	0.53	0.50	32919
weighted avg	0.55	0.53	0.50	32919

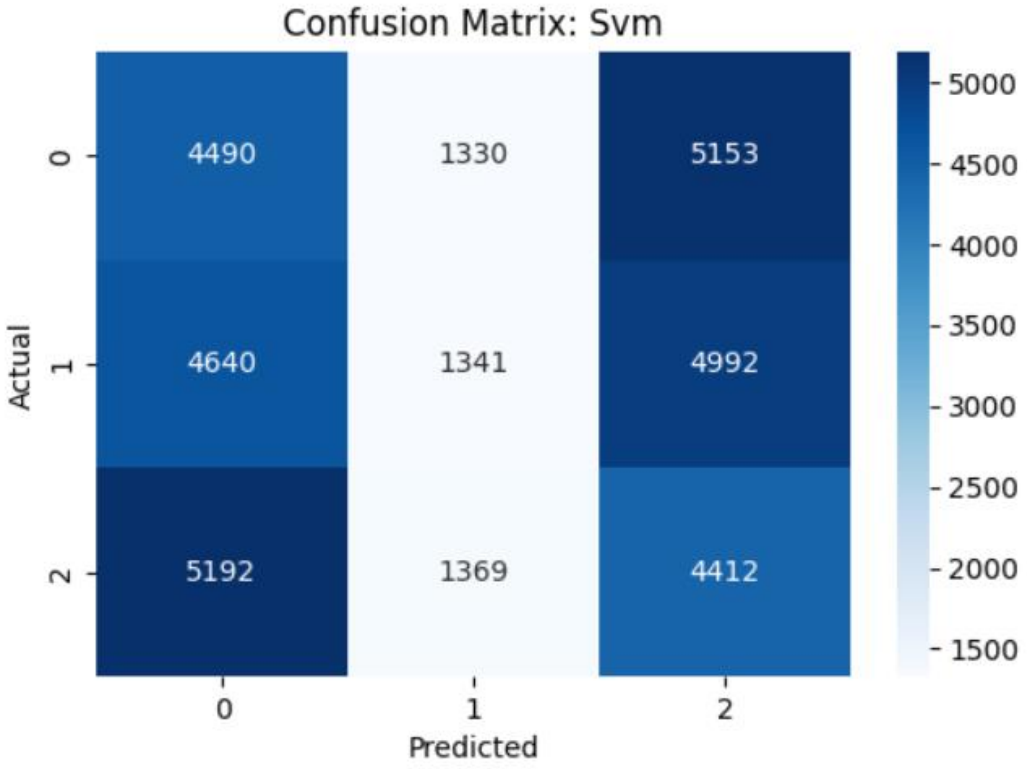
- ROC-AUC Score (macro-averaged, multi-class):

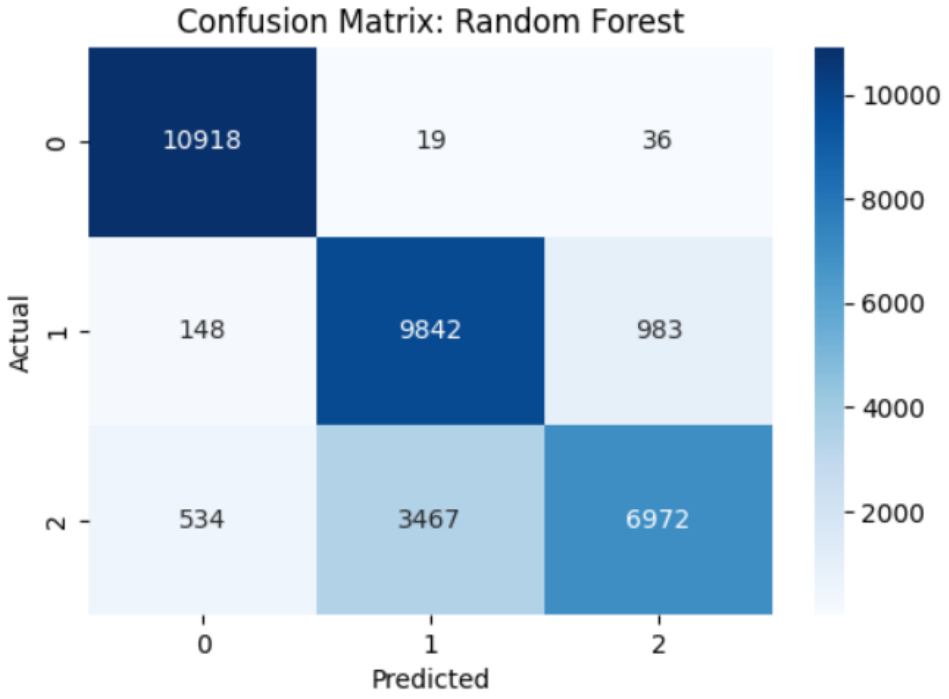
```
ROC-AUC (XGB):  0.74
ROC-AUC (Logistic Regression): 0.61
ROC-AUC (Random Forest): 0.97
```

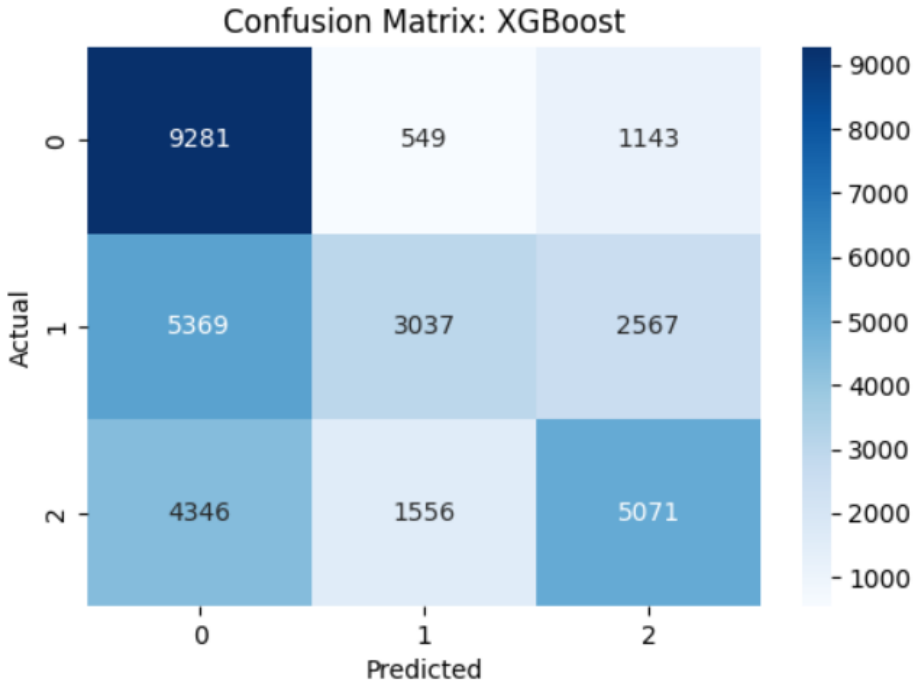
6.Confusion Matrix Visualization:

- Plots for all models to visualize performance across classes.




Confusion Matrix: Svm
4490
1330
5153
4640
1341
4992
5192
1369
4412
Actual
Predicted
0
1
2
5000
4500
4000
3500
3000
2500
2000
1500


Confusion Matrix: Random Forest
10918
19
36
148
9842
983
534
3467
6972
Actual
Predicted
0
1
2
10000
8000
6000
4000
2000



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.