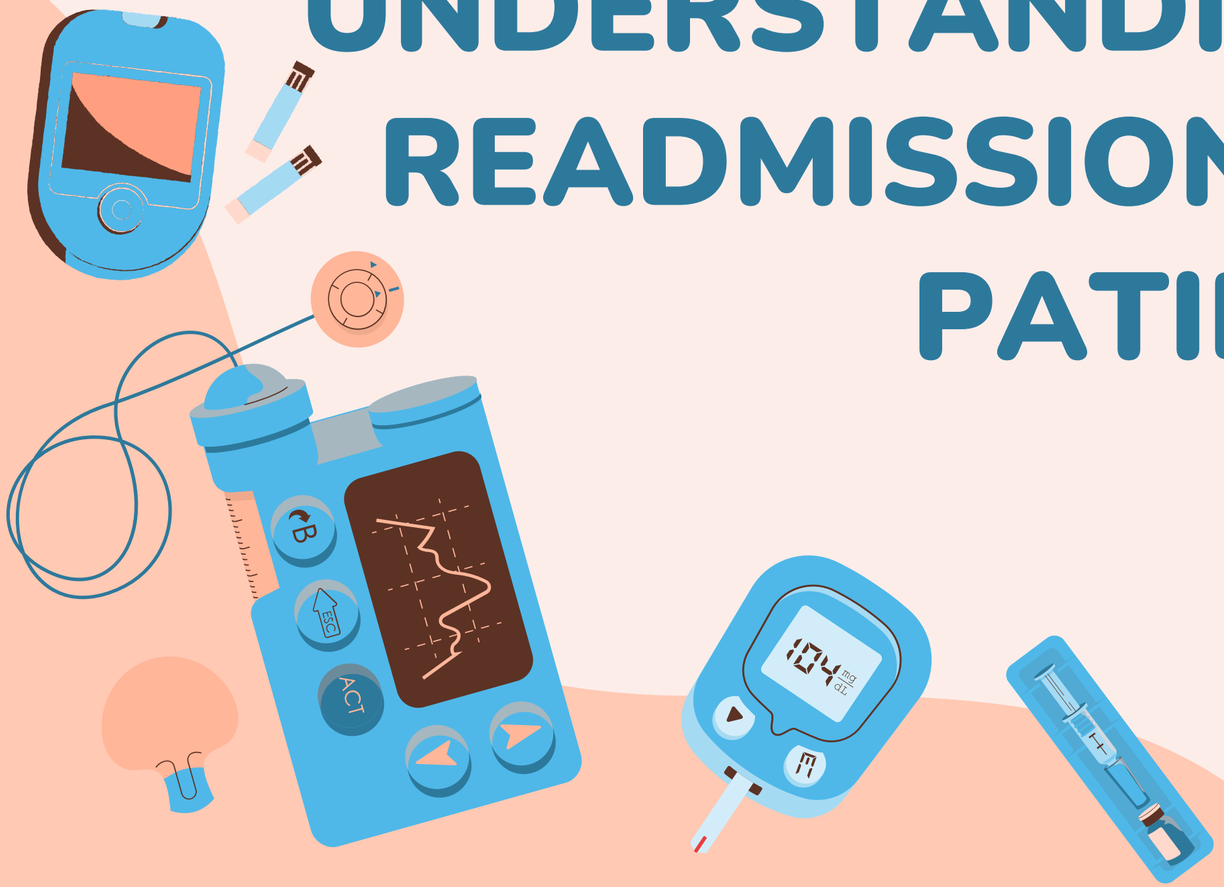
UNDERSTANDING THE EARLY READMISSION IN DIABETIC PATIENTS



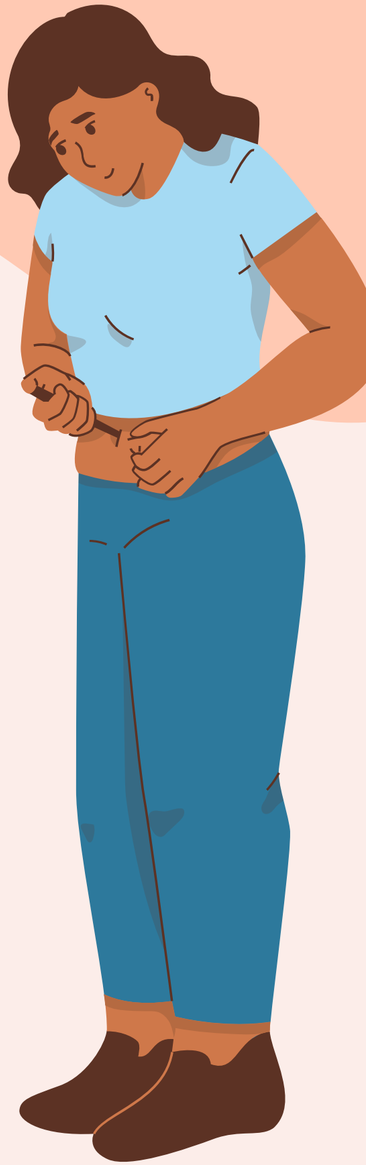
Christelle LE VAN KHANH
Anne-Solène IZARD
Lily LA MORELLA

TABLE
of contents

01. INTRODUCTION

The dataset represents ten years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. Each row concerns hospital records of patients diagnosed with diabetes, who underwent laboratory, medications, and stayed up to 14 days.

GOAL : TO DETERMINE EARLY READMISSION OF DIABETIC PATIENTS WITHIN 30 DAYS OF DISCHARGE



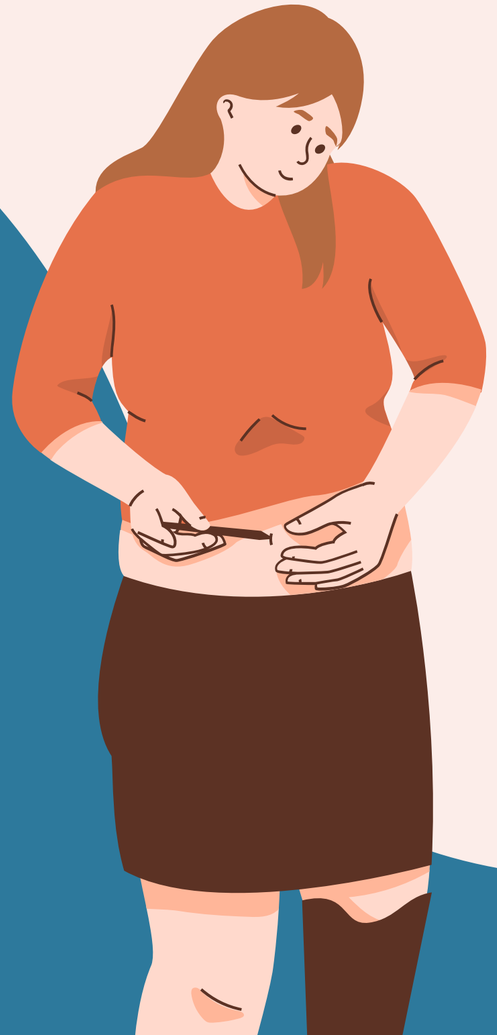
Patient Care :
Proper care reduces the likelihood of readmission, ensuring patients receive necessary treatments and support.

Better Clinical Outcomes :
Effective diabetes management during hospitalization improves patient health and prevents complications.

Managing Costs :
Addressing early readmissions helps control healthcare costs associated with managing complications.

Reducing Mortality :
Improved diabetes care reduces the risk of mortality, emphasizing the need for comprehensive management.

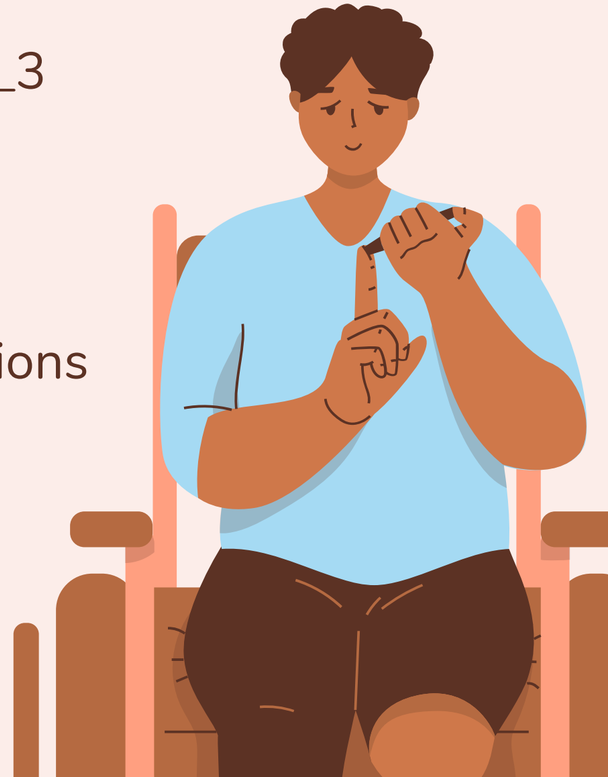
02. DATASET DETAILS

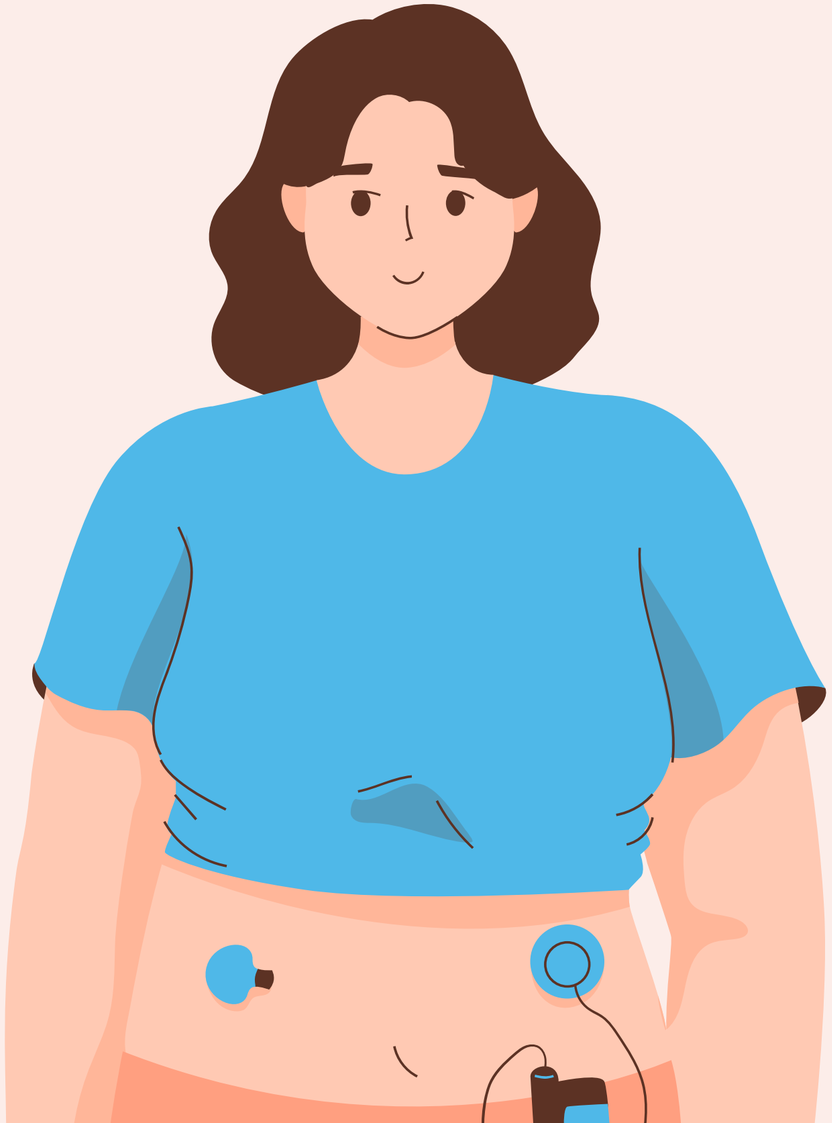


The dataset includes 50+ features representing patient and hospital outcomes. Data was extracted from inpatient encounters with diabetes diagnoses, lasting 1 to 14 days, involving laboratory tests and medication administration.

Here is the list of the columns before any cleaning :

- encounter_id
- patient_nbr
- race
- gender
- age
- weight
- admission_type_id
- discharge_disposition_id
- admission_source_id
- time_in_hospital
- payer_code
- medical_specialty
- num_lab_procedures
- num_procedures
- num_medications
- number_outpatient
- number_emergency
- number_inpatient
- diag_1, diag_2, diag_3
- number_diagnoses
- max_glu_serum
- A1Cresult
- 22 different medications
- change
- diabetesMed
- readmitted





03. DATA PREPROCESSING

DATA CLEANING

1

Replace the “?” by NaN
Looking for null values and dropping columns that have more than 35% of null values

2

Verifying the unicity of values in columns, we decided to drop columns that have more than 98% of the same value (irrelevant)

3

Mapping Variables

4

Replace the NaN in the race column by the most frequent race

DATA CLEANING : FOR CLASSIFICATION

1

We classify the different values of each column and put them in a logical order to have some link between close values

2

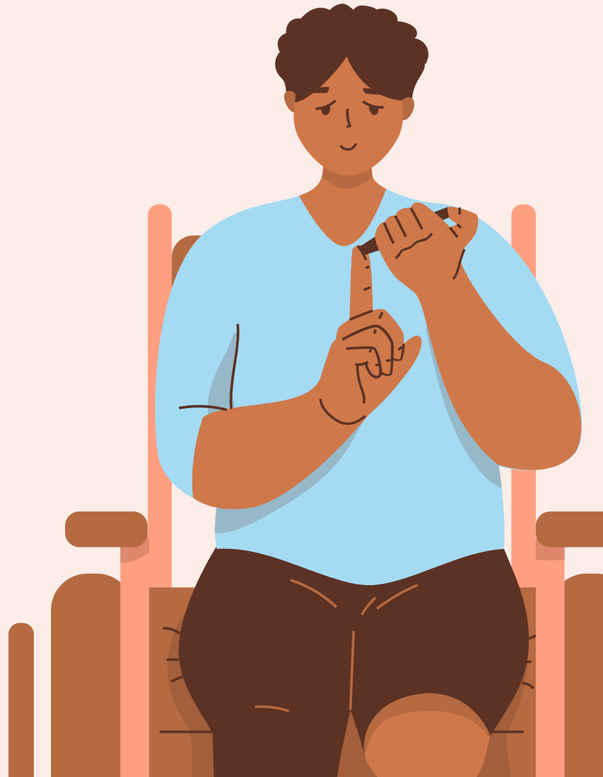
We have to transform all our data into numerical data

According to our precedent observations we dropped :

- weight
- payer_code
- medical_specialty

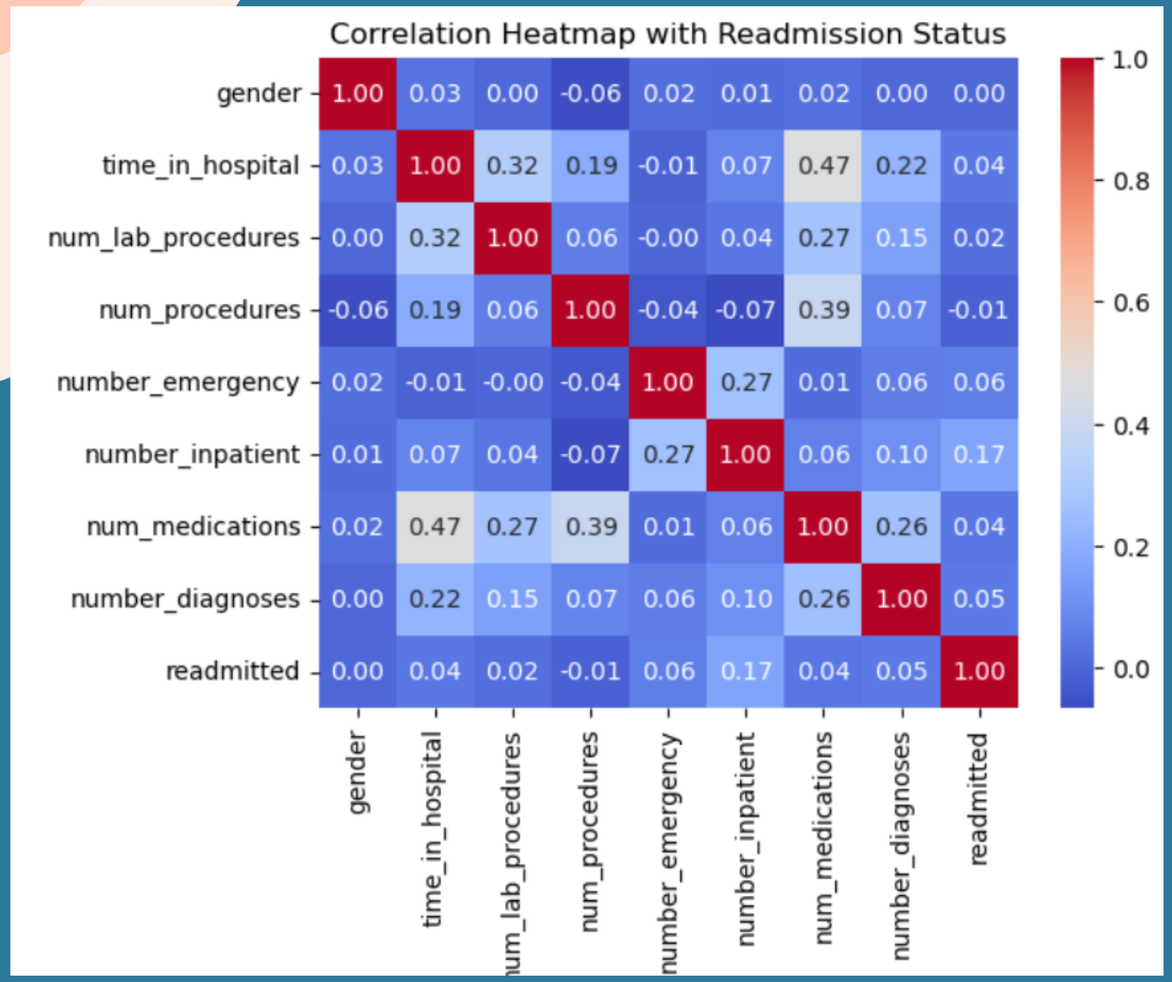
- Medications below :

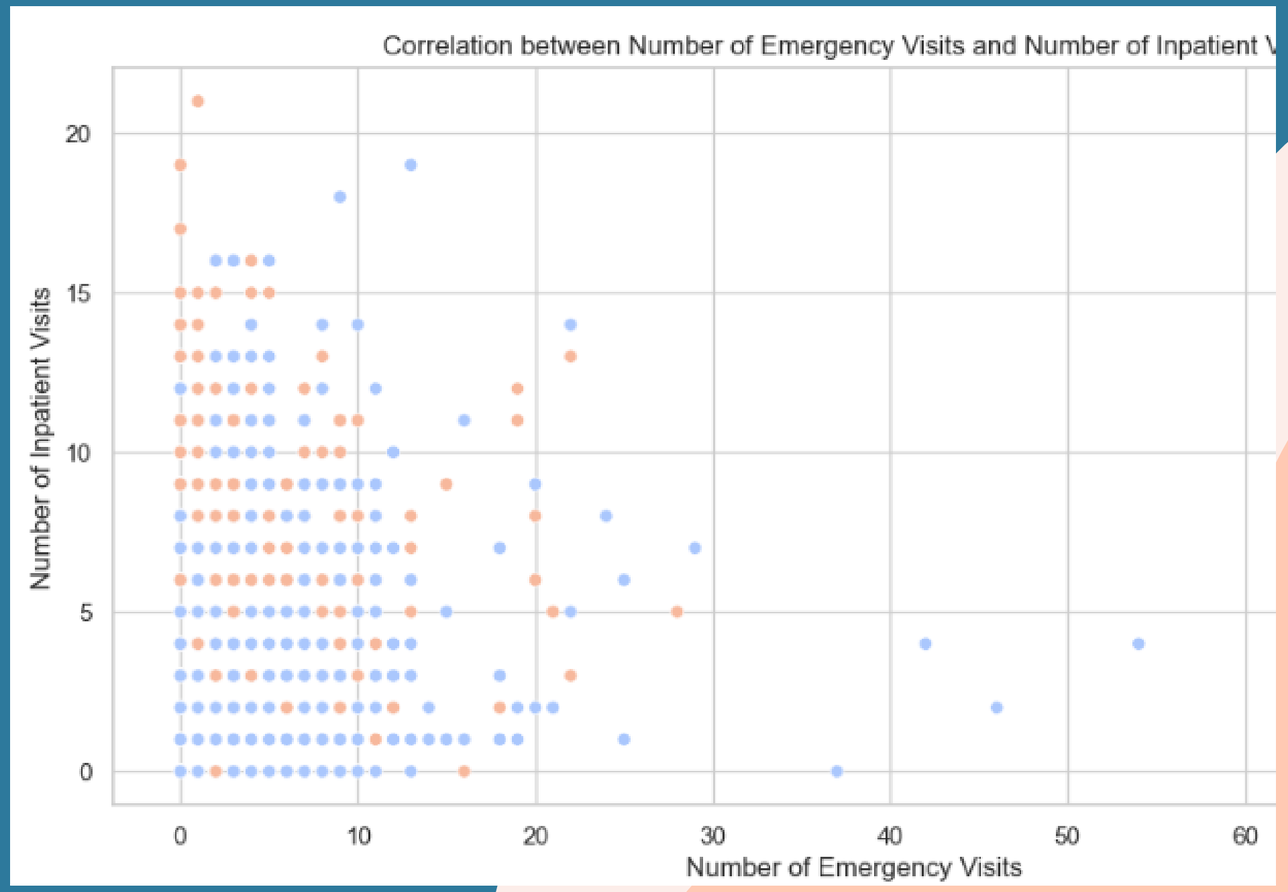
examide', 'citoglipton','acetohexamide', 'troglitazone', 'glipizide-metformin', 'glimepiride-pioglitazone','metformin-rosiglitazone', 'metformin-pioglitazone', 'tolbutamide','repaglinide', 'nateglinide', 'chlorpropamide', 'acarbose', 'miglitol', 'tolazamide','glyburide-metformin'

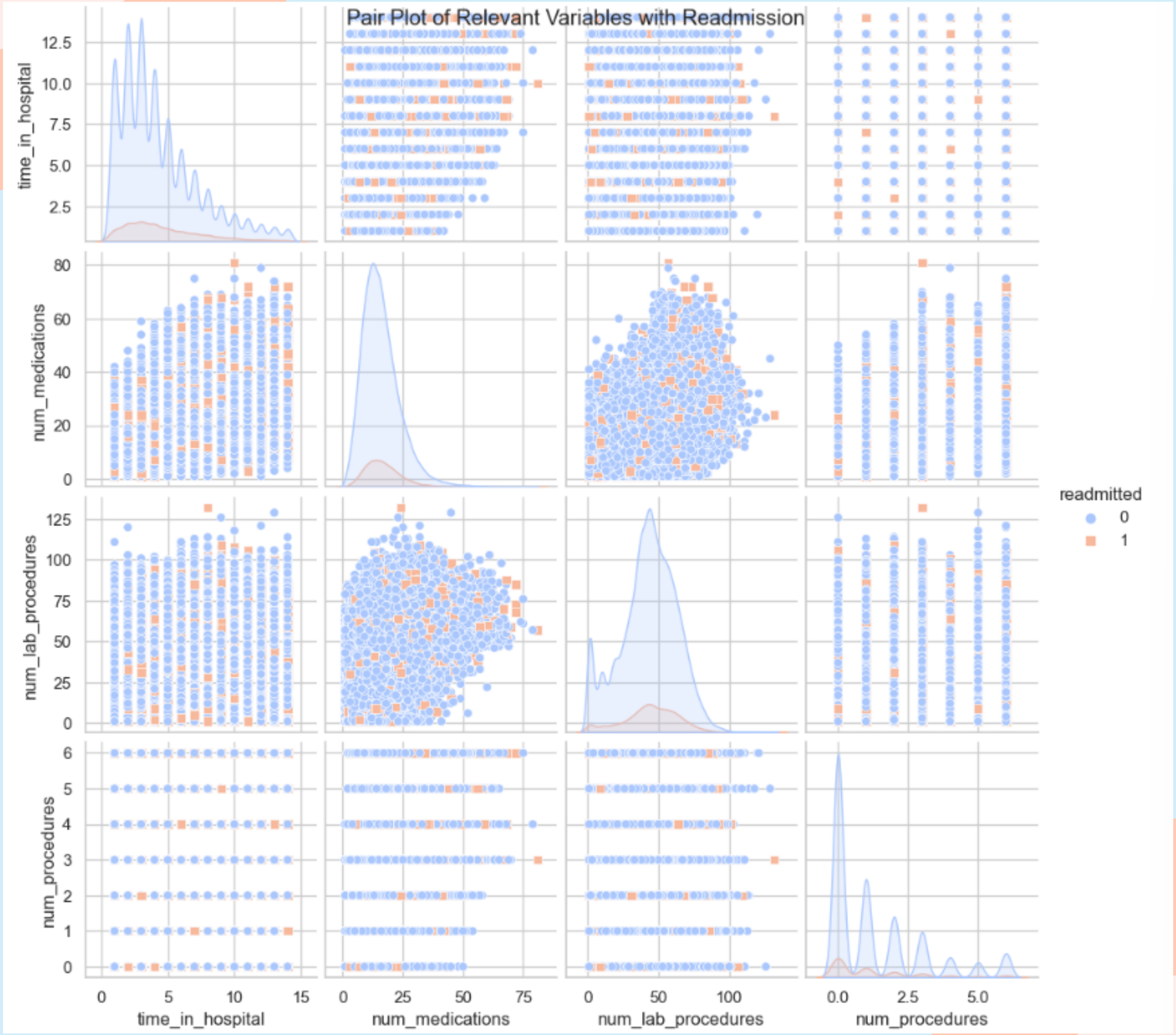


04. DATA VISUALISATIONS

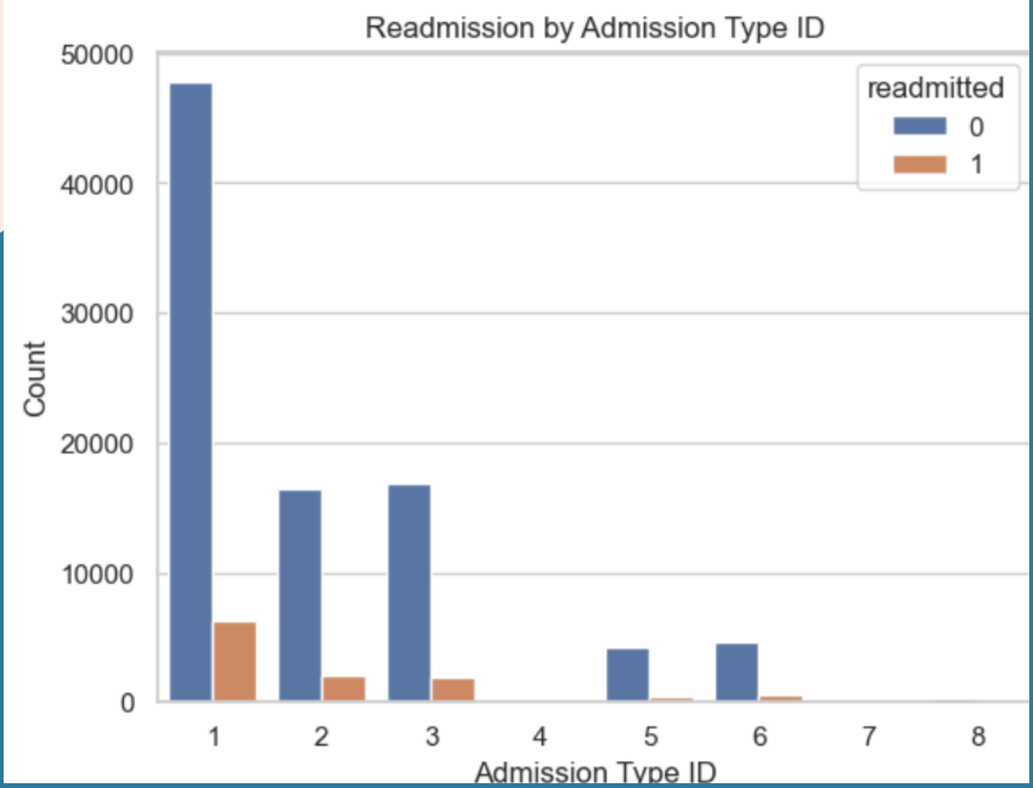
CORRELATION STUDY

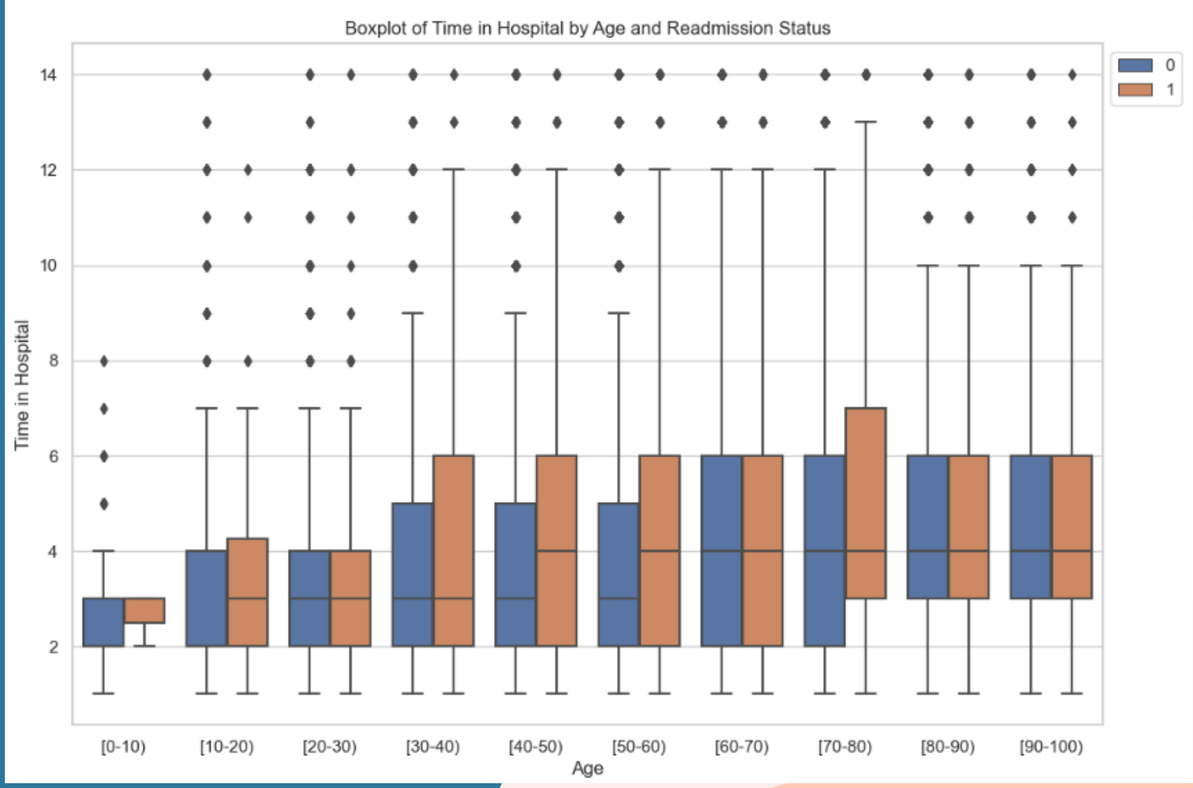


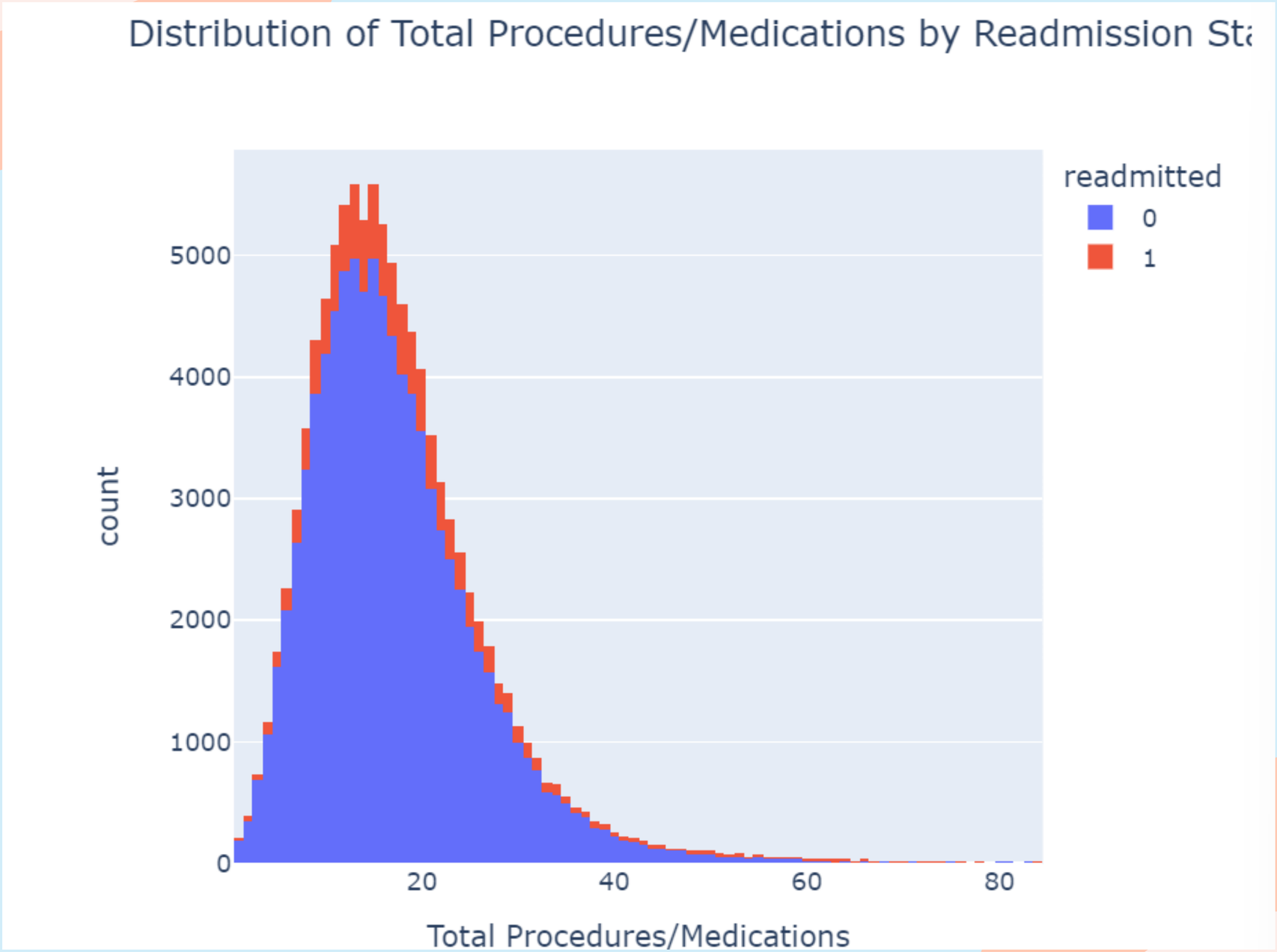



Pair Plot of Relevant Variables with Readmission
time_in_hospital
num_medications
num_lab_procedures
num_procedures
readmitted
0
1

TARGET VARIABLE : READMITTED






Distribution of Total Procedures/Medications by Readmission Sta
readmitted
0
1
count
5000
4000
3000
2000
1000
0
20
40
60
80
Total Procedures/Medications

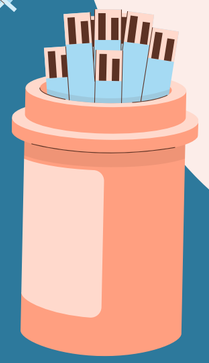
05. MACHINE LEARNING MODELS

About the dataset :

We have a repartition of 0 and 1 in the column “readmitted” that is imbalanced.

90% against **10%**

That makes some metrics unusables like the accuracy.



Usefull metrics :

- Precision
- Recall
- F1 score
- Confusion matrix

Resampeling the dataset :

We used oversampeling to try to balance the dataset.

It lead to overfiting and didn’t give good results

Possible Classification Models :

- Logistic Regression
- K Nearest Neighbors
- Gradient Boosting Classifier
- Random Forest Classifier
- 👍 **Balanced Random Forest Classifier**
- Balanced Bagging Classifier

THANK YOU!

Thank you so much for watching our presentation !