DIA 5

GABRIEL CHAIX | MALO LEROY | SEBASTIEN MUNIER

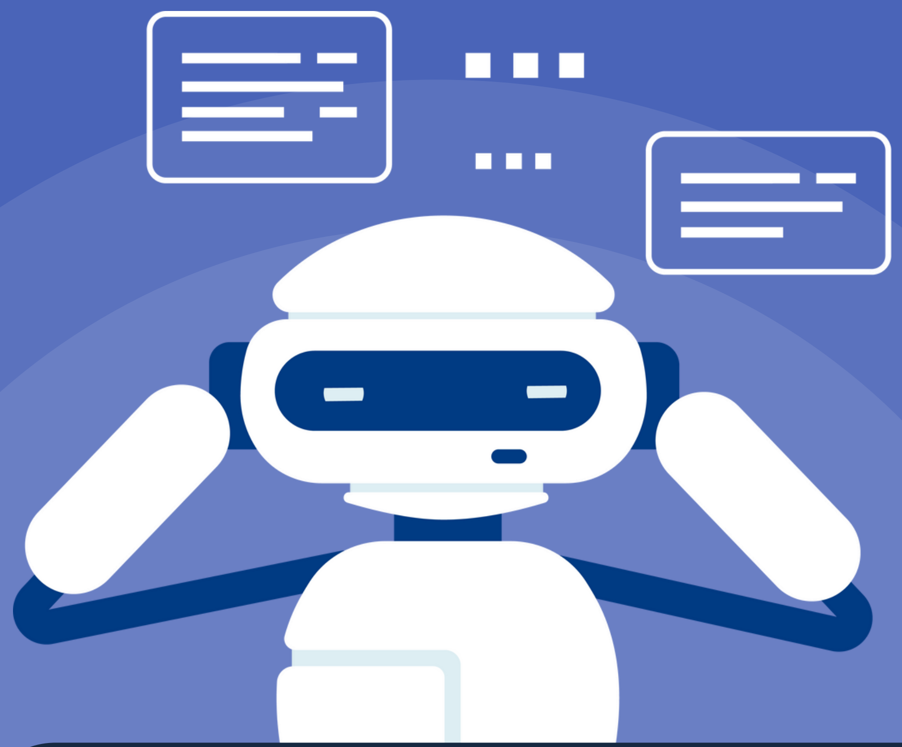
ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

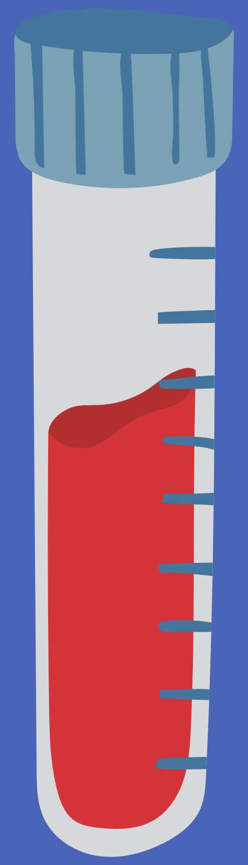
Python for Data Analysis

[PROJECT] Diabetes in US Hospitals

S1 2023-2024 | **ESILV**

Diabetes Insight





What cost ?

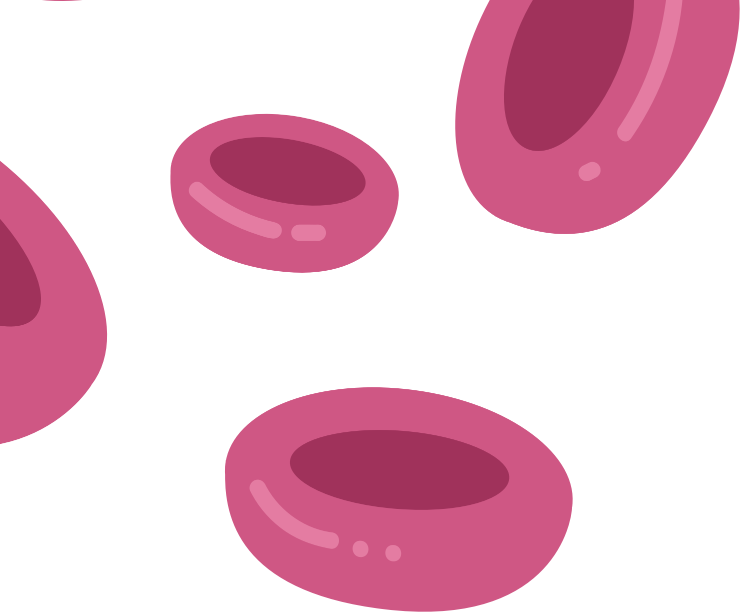
In 2022, the total cost of diagnosed diabetes in the United States was estimated at **412.9 billion S**

Diagnosed ?

In 2019, about **11.3% of the U.S. population** were living with diabetes, with two-thirds diagnosed and one-third undiagnosed

Prediabetes

Prediabetes, characterized by blood glucose levels above normal but not high enough for a diabetes diagnosis, affects over one-third of adults and is a key risk factor for type 2 diabetes



Content Outline

01 Overview of the dataset

02 Cleaning and pre-processing

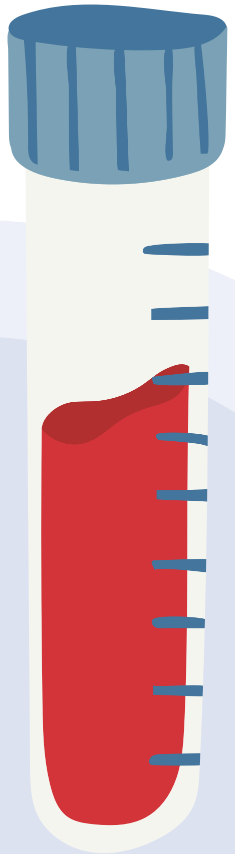
03 Data visualization for understanding

04 Data modeling and fine-tuning

05 Deployment uisng flask

OVERVIEW OF THE DATASET

Diabetes 130 US hospitals



The data set represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes.

Cleaning and pre-processing

What was removed

Many missing values :
Weight, payer_code, medical_specialty
Only-one-value columns
Examide, citoglipton, metformin-rosiglitazone
Any-null-value rows

numchange

We have sum time each medecine was given. High count mean they were not cured with de proper medecin.

number_services

We chose to treat number_outpatient, number_emergency and number_inpatient in a unique column to quantify global usage of hospital services by the patient.

readmitted

To reduce our problem to a binary classification, we combined the readmission after 30 days and no readmission into a single category

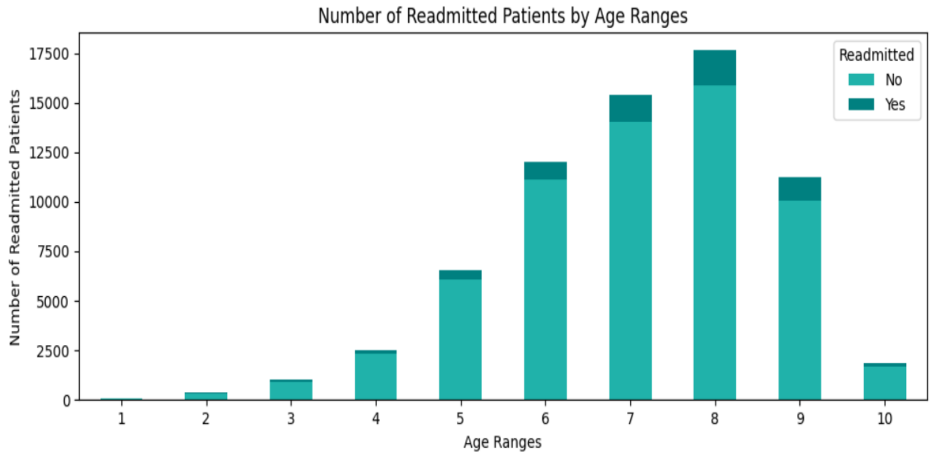
What was removed?

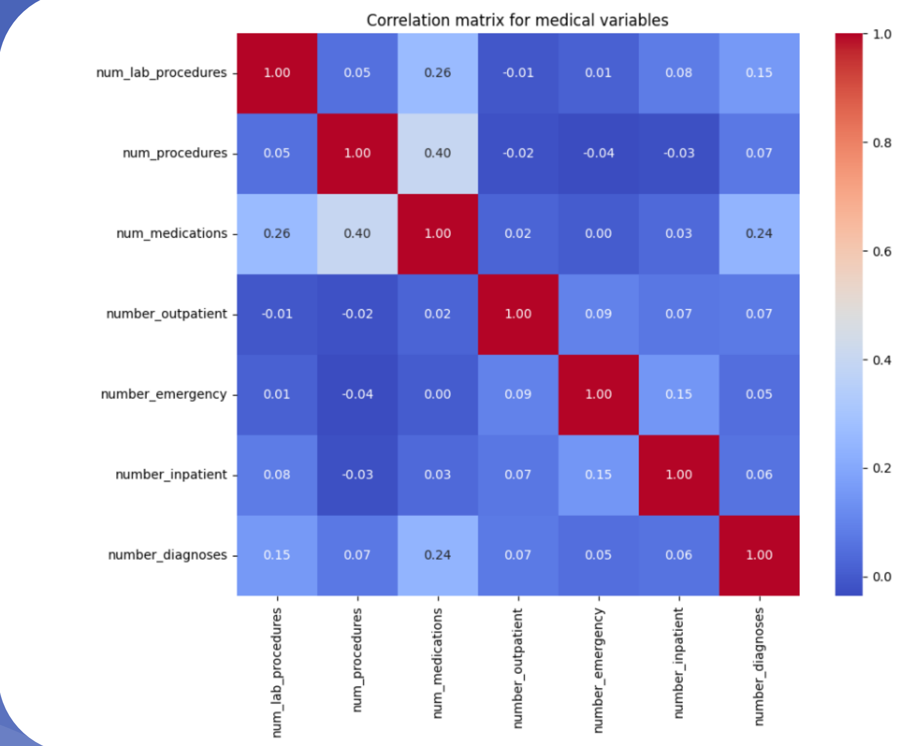
New variables explanation

Data Visualization

Plot 1 : Number of readmitted people by age ranges

Most of the diabetic people are aged between 50 and 80 years old. The proportion of readmission is close for each range.


Number of Readmitted Patients by Age Ranges


Correlation matrix for medical variables

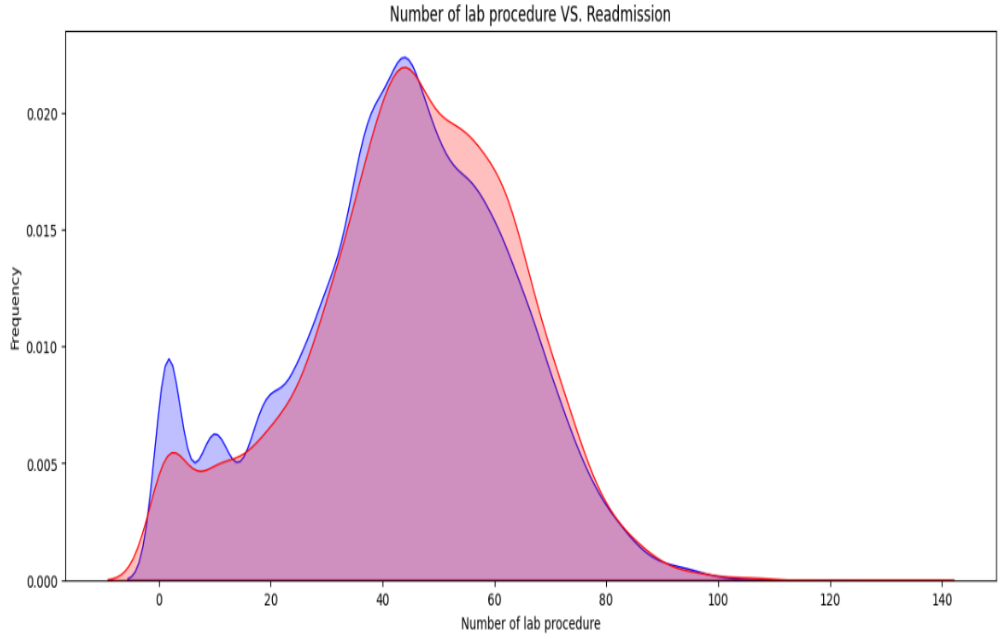
	num_lab_procedures	num_procedures	num_medications	number_outpatient	number_emergency	number_inpatient	number_diagnoses
num_lab_procedures	1.00	0.05	0.26	-0.01	0.01	0.08	0.15
num_procedures	0.05	1.00	0.40	-0.02	-0.04	-0.03	0.07
num_medications	0.26	0.40	1.00	0.02	0.00	0.03	0.24
number_outpatient	-0.01	-0.02	0.02	1.00	0.09	0.07	0.07
number_emergency	0.01	-0.04	0.00	0.09	1.00	0.15	0.05
number_inpatient	0.08	-0.03	0.03	0.07	0.15	1.00	0.06
number_diagnoses	0.15	0.07	0.24	0.07	0.05	0.06	1.00

Plot 2 : Correlation matrix for medical variables

The variables don't represent any correlation between each other. We found the same results for the whole dataset. This enlights the importance of machine learning model for our analysis.

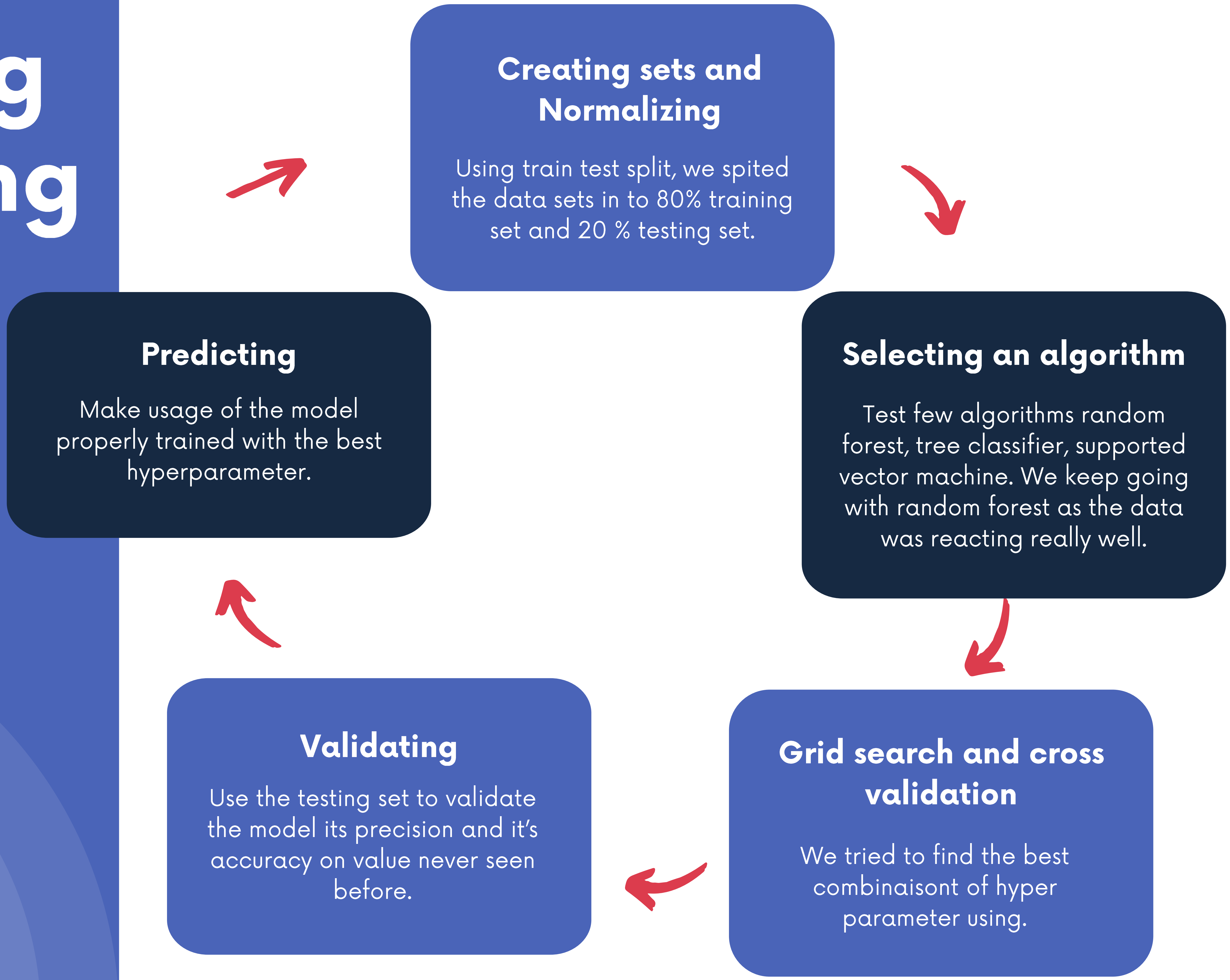
Plot 3 : Frequency of lab procedures vs readmission

The graph presents a very close correlation between readmitted and non-readmitted people, depending on the frequency of lab procedures.


Number of lab procedure VS. Readmission

Data modeling and fine-tuning

A multi-step choice making process to find what's adapted to our data.



Creating sets and Normalizing

Using train test split, we spited the data sets in to 80% training set and 20 % testing set.

Selecting an algorithm

Test few algorithms random forest, tree classifier, supported vector machine. We keep going with random forest as the data was reacting really well.

Grid search and cross validation

We tried to find the best combinaisont of hyper parameter using.

Validating

Use the testing set to validate the model its precision and it's accuracy on value never seen before.

Predicting

Make usage of the model properly trained with the best hyperparameter.

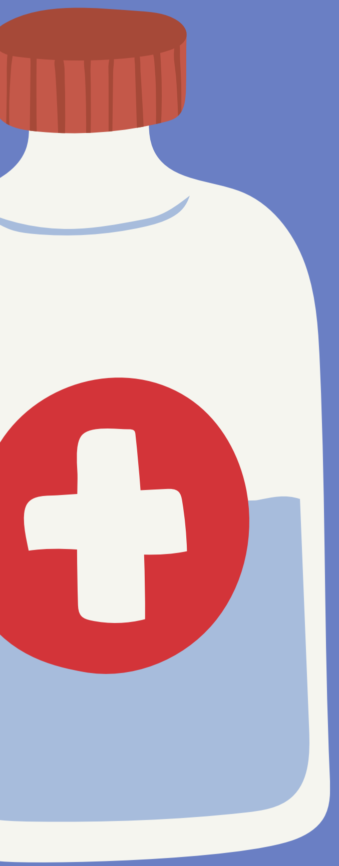
Deployment

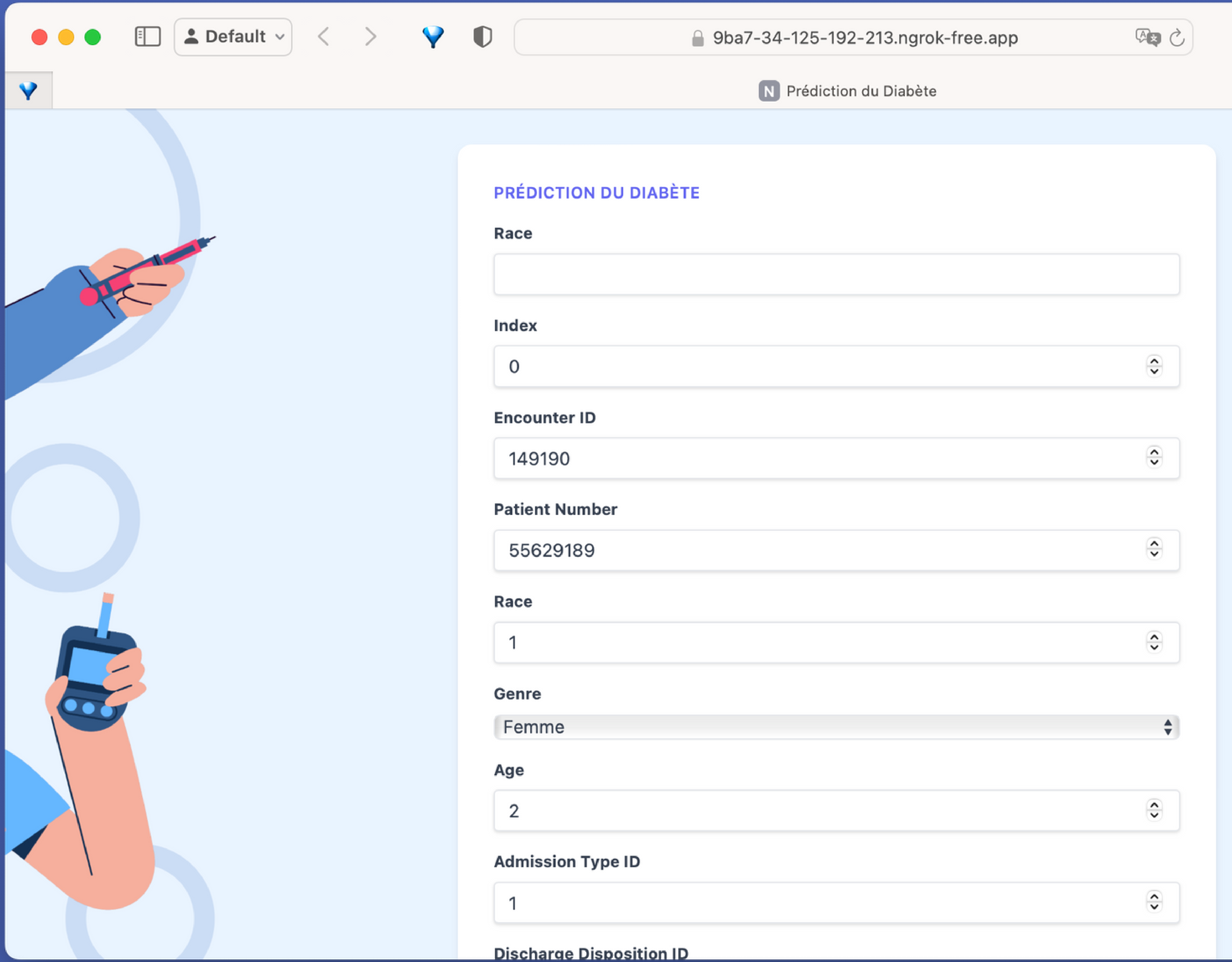
Pyngork and Flask from colab

Using the model we’ve trained and saved in our drive, we can deploy and host a web page directly from the jypyter/colab notebook. We have combine the usage of flask and pyngork to use our trained model to predict wether or not a person have chances of readmition by answering a forms.

You can also acces the forms at :

https://diabetes.gab.cx





PRÉDICTION DU DIABÈTE

Race

Index
0

Encounter ID
149190

Patient Number
55629189

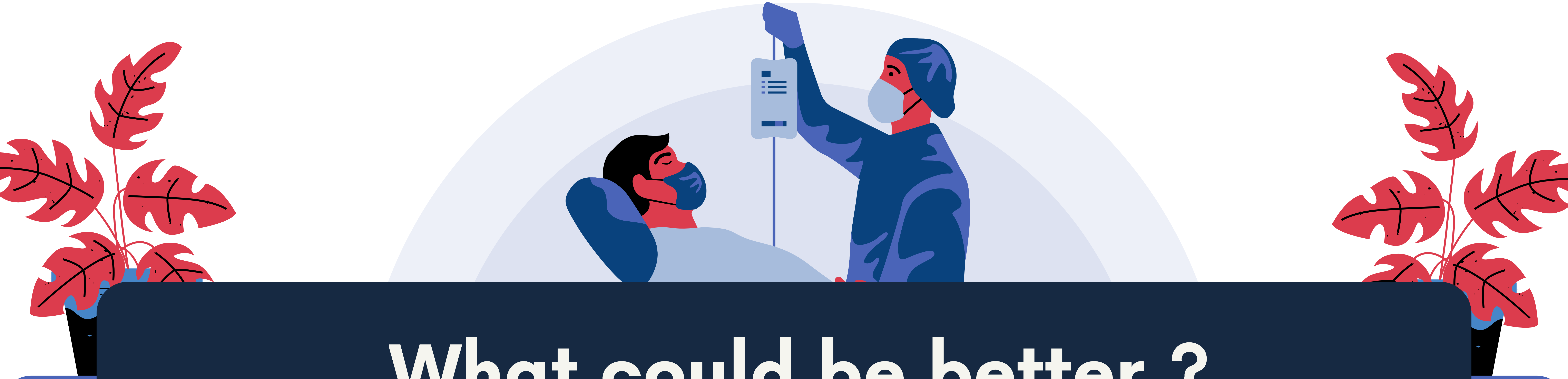
Race
1

Genre
Femme

Age
2

Admission Type ID
1

Discharge Disposition ID



What could be better ?

Incorporating patient weight data into our model can enhance accuracy in identifying diabetes risk, due to weight being a significant risk factor. However, the challenge lies in the significant number of missing values in our dataset's weight data

Thanks for reading !

Questions or comments?

Email Address

malo.leroy@edu.devinci.fr
sebastien.mugnier@edu.devinci.fr
gabriel.chaix@edu.devinci.fr

Website

https://github.com/malolry/Project-Data-Analyst