Final Project

Python For Data Analysis

NOYELLE Amélie KEENER Jeremy PENIGUEL Josselin - DIA1
Supervised by Mr SABRY

January 2022

Project Description

- To finish our semester of Python for data Analysis, we have been working on a project that we will present in these pages. We were given a dataset and we had to process and analyze the data and use it to make predictions using machine learning.

- Our dataset gathers information from 101767 admissions in 130 US hospitals between 1999 and 2008. All patients had diabetes and stayed in the hospital between 1 and 14 days. A total of 50 variables for each admission are provided.

- We decided to predict if a patient according to his information will be readmit at the hospital.

DataSet Description

Variables

encounter_id	int64
patient_nbr	int64

The first 2 variables allow us to identify the patient and the admission to the hospital.

race	object
gender	object
age	object
weight	object

The above variables give us basic information about ethnicity, gender, age and weight.

admission_type_id	int64
discharge_disposition_id	int64
admission_source_id	int64
time_in_hospital	int64
payer_code	object
medical_specialty	object
num_lab_procedures	int64
num_procedures	int64
num_medications	int64
number_outpatient	int64
number_emergency	int64
number_inpatient	int64

Then we have some information related to the proper functioning of the hospital, such as the type of payment, the time spent in the hospital, the department in which the patient was admitted.

diag_1	object
diag_2	object
diag_3	object
number_diagnoses	int64

We then have the diagnosis made by the doctors when the patient arrives (the 3rd opinion is used if the first 2 are not the same)

max_glu_serum	object
A1Cresult	object
metformin	object
repaglinide	object
nateglinide	object
chlorpropamide	object
glimepiride	object
acetohexamide	object
glipizide	object
glyburide	object
tolbutamide	object
pioglitazone	object
rosiglitazone	object
acarbose	object
miglitol	object
troglitazone	object
tolazamide	object
examide	object
citoglipton	object
insulin	object
glyburide-metformin	object
glipizide-metformin	object
glimepiride-pioglitazone	object
metformin-rosiglitazone	object
metformin-pioglitazone	object
change	object

We have above a long list of analysis results

diabetesMed	object
readmitted	object

The last two variables tell us if the patient has received any diabetes-related medication and if he/she has been readmitted and since when.

NOYELLE Amélie –KEENER Jeremy – PENIGUEL Josselin

DataSet Cleaning

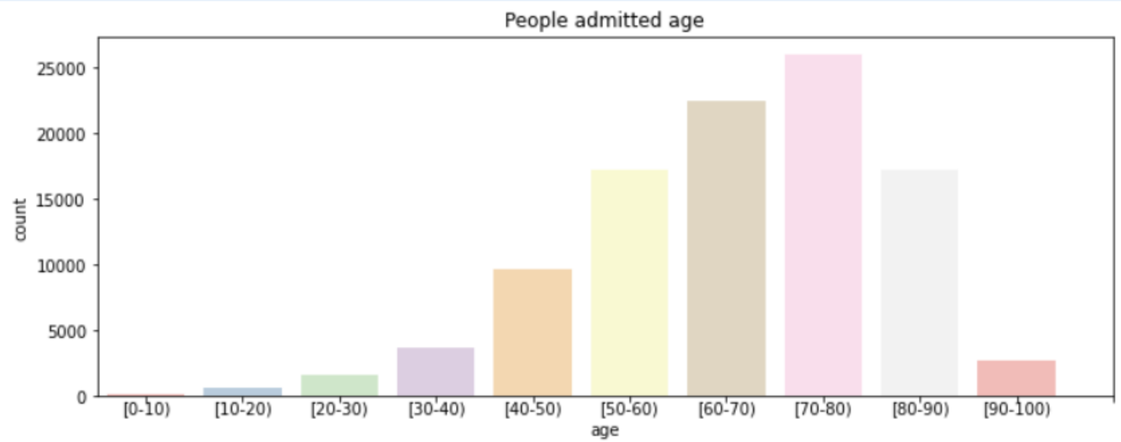
- Replace all the “?” and “Unknown” cells by a NaN one.
- Change all the qualitative variables into quantitative ones. We create dictionary to get all the information ordered.
- Calculate the variance of all the variable
- Choose which column to drop : here we drop “Weight”, “payer_code”, “medical_speciality” because they lack too much information.

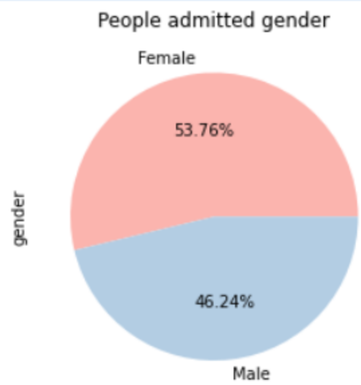
column_name	percent_missing
weight	96.858479
payer_code	39.557416
medical_specialty	49.082208

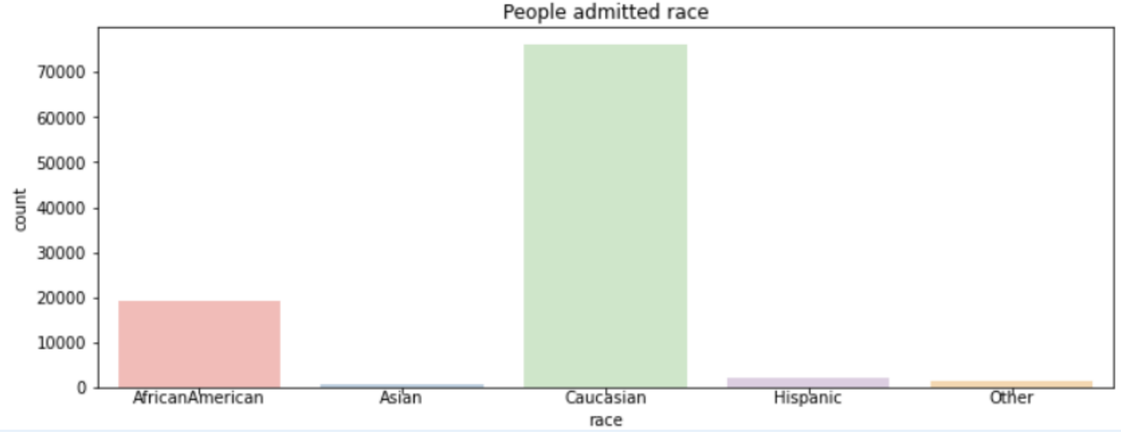
- We also drop all variable that have a variance under 10e-1

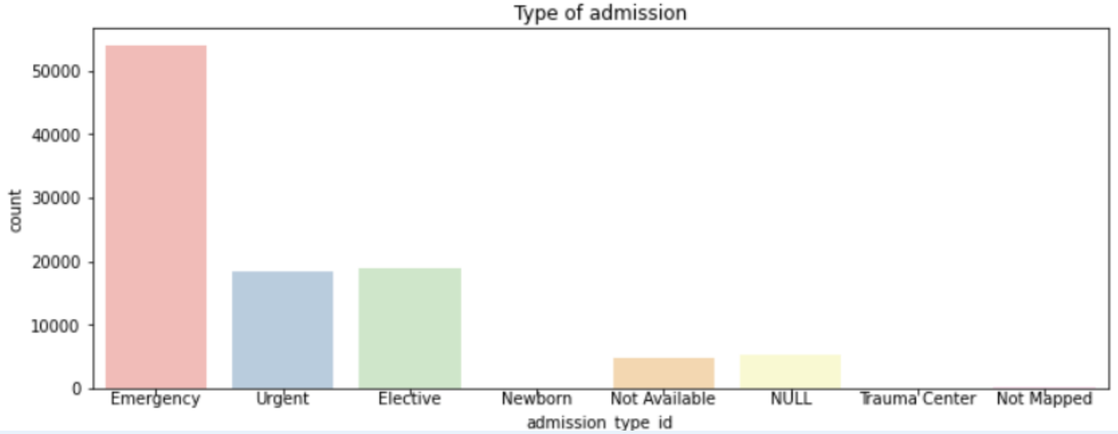
Data Visualization

Global data visualization





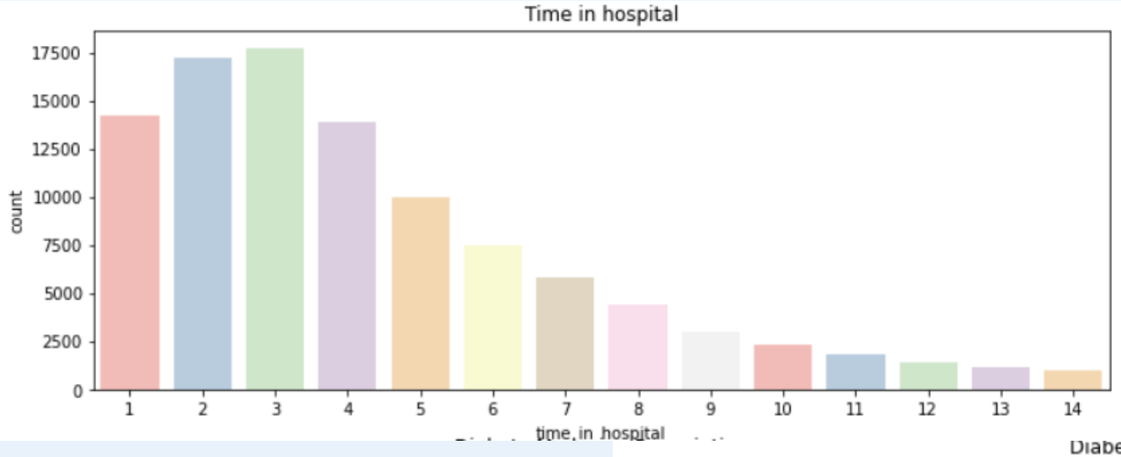


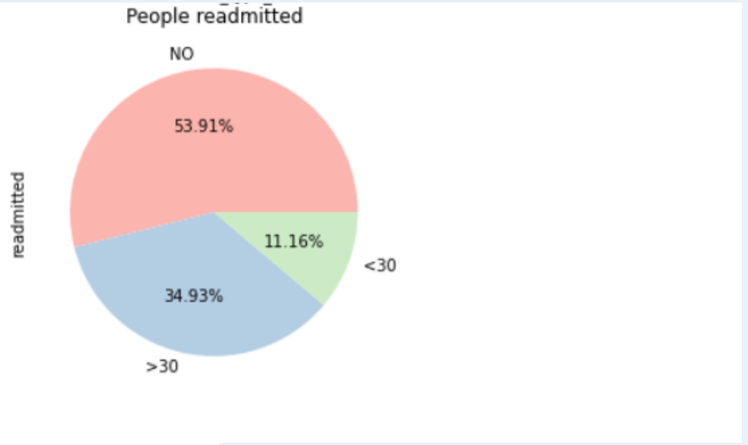


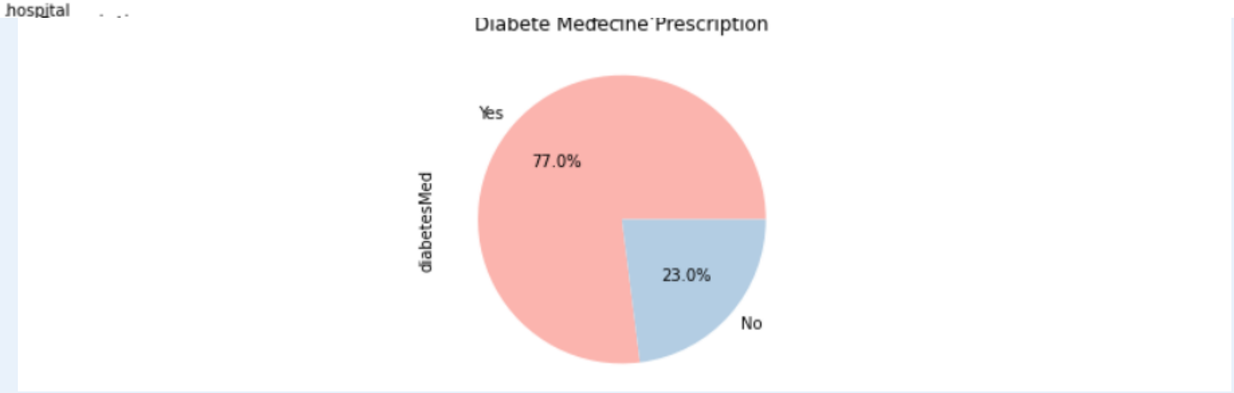
NOYELLE Amélie –KEENER Jeremy – PENIGUEL Josselin

Data Visualization

Global data visualization

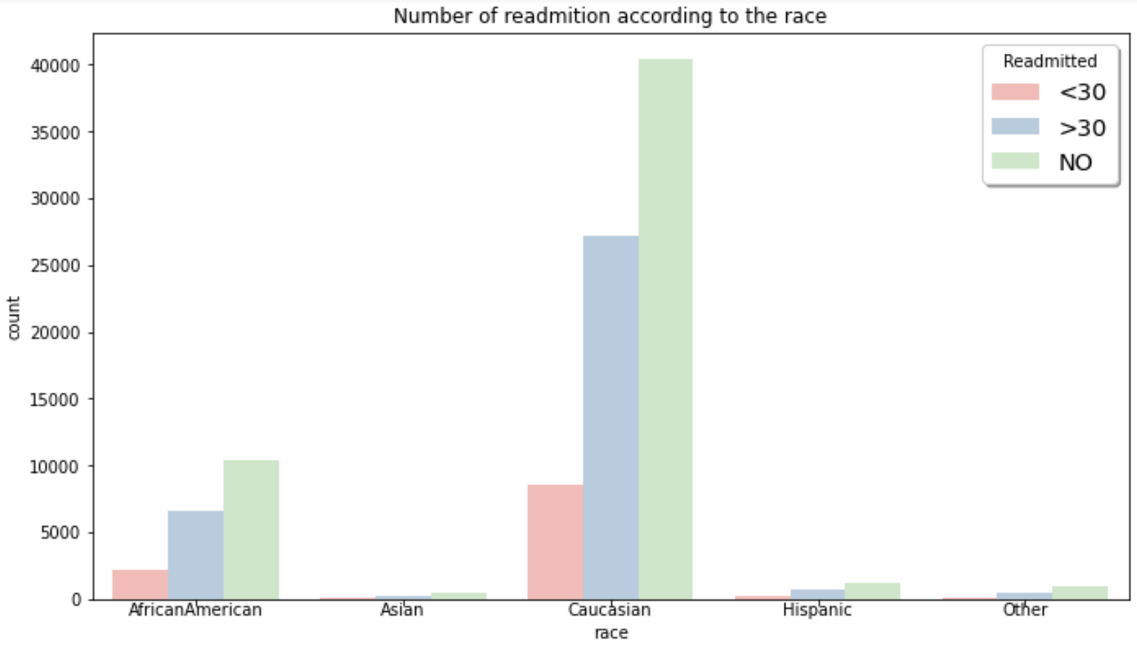


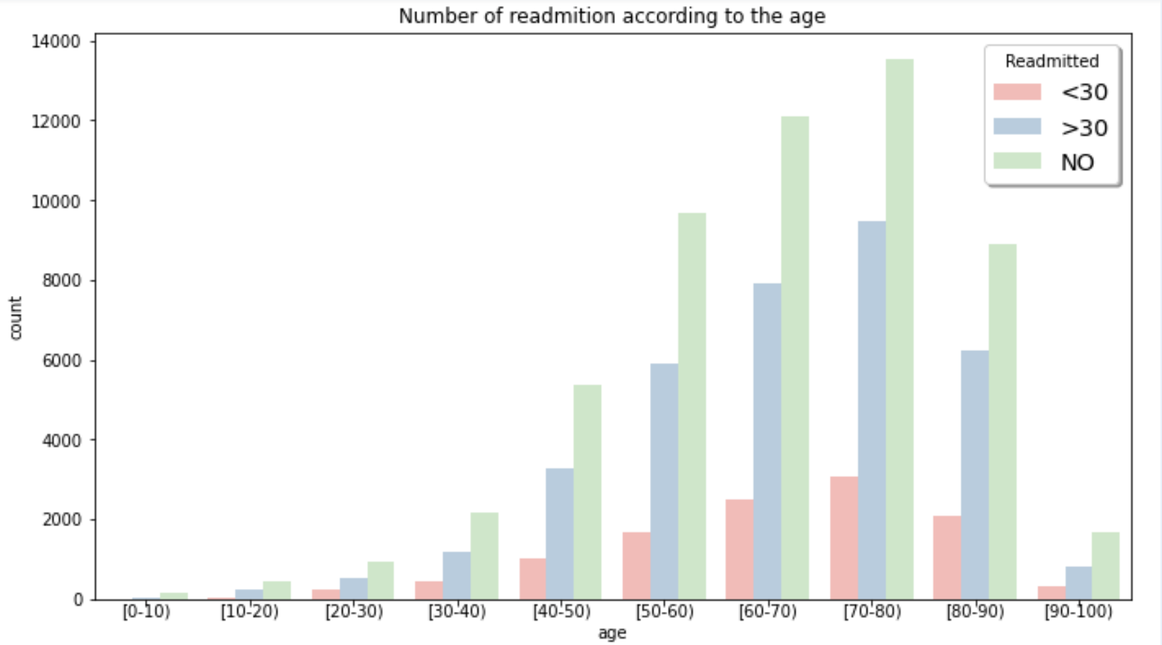




Data Visualization

Visualization according to the readmission

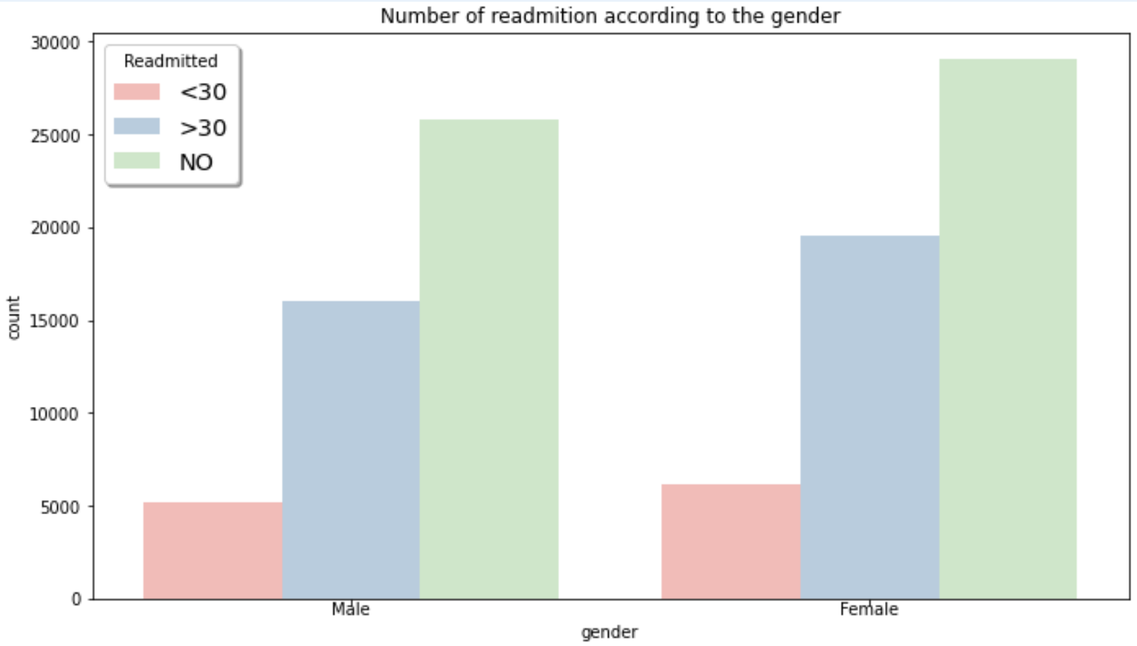


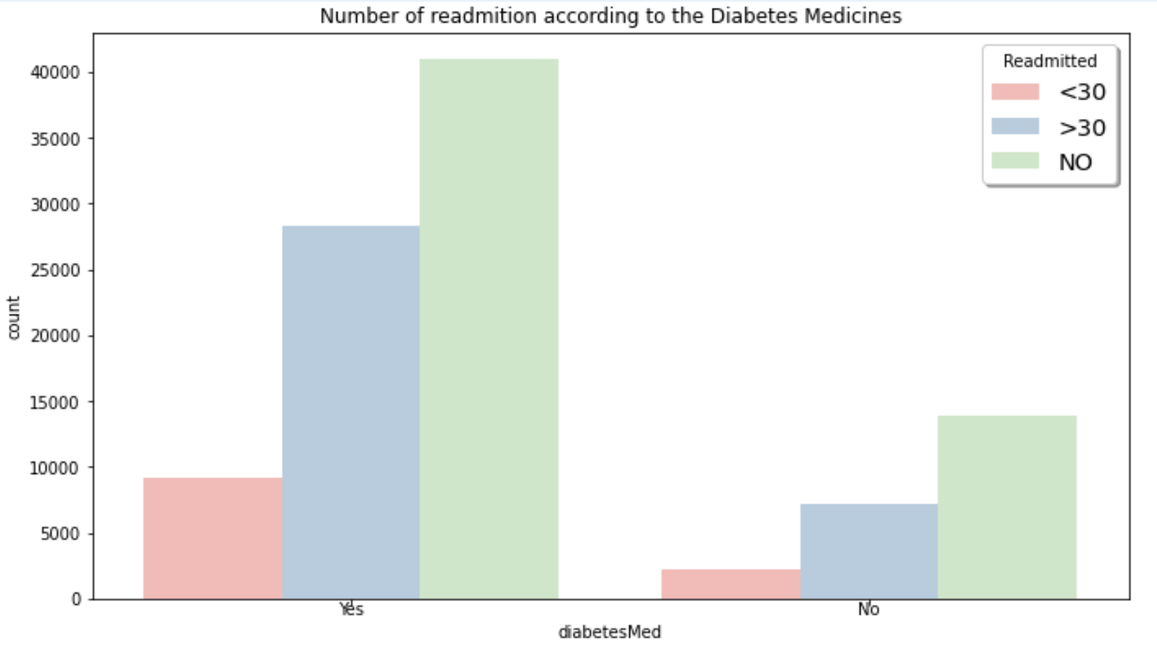


NOYELLE Amélie –KEENER Jeremy – PENIGUEL Josselin

Data Visualization

Visualization according to the readmission



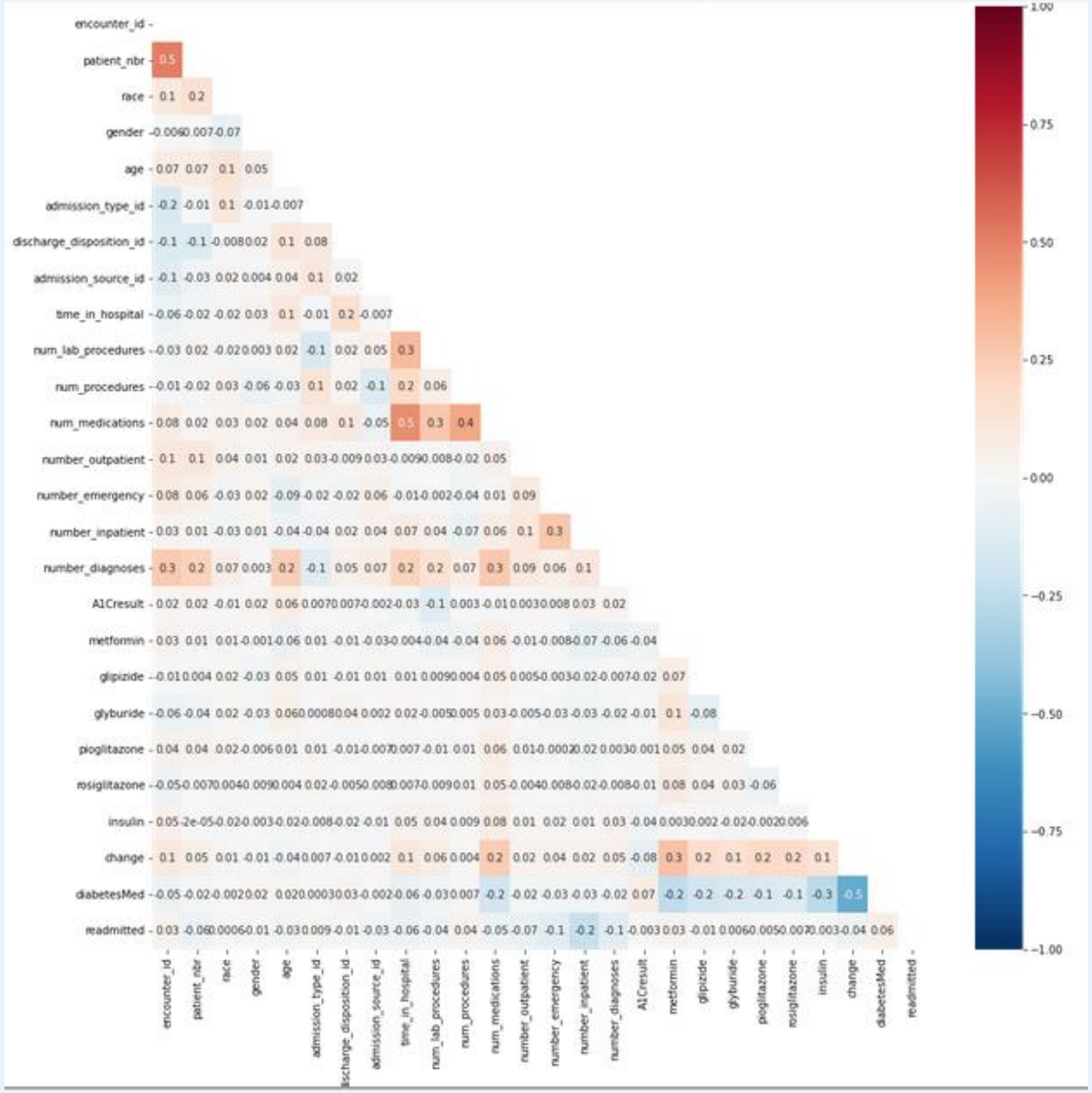


Machine Learning

Correlation Matrix

- The first step of our Machine Learning process was to make a correlation matrix in order to have a visual on the variables that are strongly correlated between them.

- We can see that number of medication and time stay in hospital are highly correlated as well as change (blood exam change or not during the stay in hospital) and DiabetesMed (if the patient had diabetes treatment) and Medication number with the procedure number.



NOYELLE Amélie –KEENER Jeremy – PENIGUEL Josselin

Machine Learning

Prepare data

- First we check that all variables are float type or int type to integrate them to our models correctly.
- We drop all the lines with 'NaN' cell to avoid errors.
- We split our dataset in two. The first part is composed only of the variable readmitted, the second of all t he other variables that we have kept except 'patient_nbr' and 'encounter_id' which will not bring us any i nformation.
- We then separate our dataset into train and test.
- Then we check that all there is no 'NaN' or infinite number in our dataset.
- And we finally scale the data

Machine Learning

Models

- We first test KNN model. We obtain 0.5083% of success. We create a confusion matrix to see where are the errors.

```
0.4916462463800401
[[ 182  630  688]
 [ 392 1876 2520]
 [ 458 2158 4563]]
0.4871909111160615
[[ 150  597  753]
 [ 399 1890 2499]
 [ 434 2224 4521]]
```

- We try to implement the model on data that are not scaled and we obtain 0.5011% of success which is almost the same as the previous result.

- Then we test Random Forest model with the RandomForestClassifier() function. We obtain 0.5316% of success, which is better than the KNN result.

```
0.5736244152372466
[[  34  620  846]
 [  22 1883 2883]
 [  10 1361 5808]]
```

- We finally test Gradient Boosting model with the GradientBoostingClassifier() function. We obtain 0.5813% of success, which is better than all the other result.

```
0.5813469963614762
[[  31  609  860]
 [  23 1699 3066]
 [   9 1071 6099]]
```

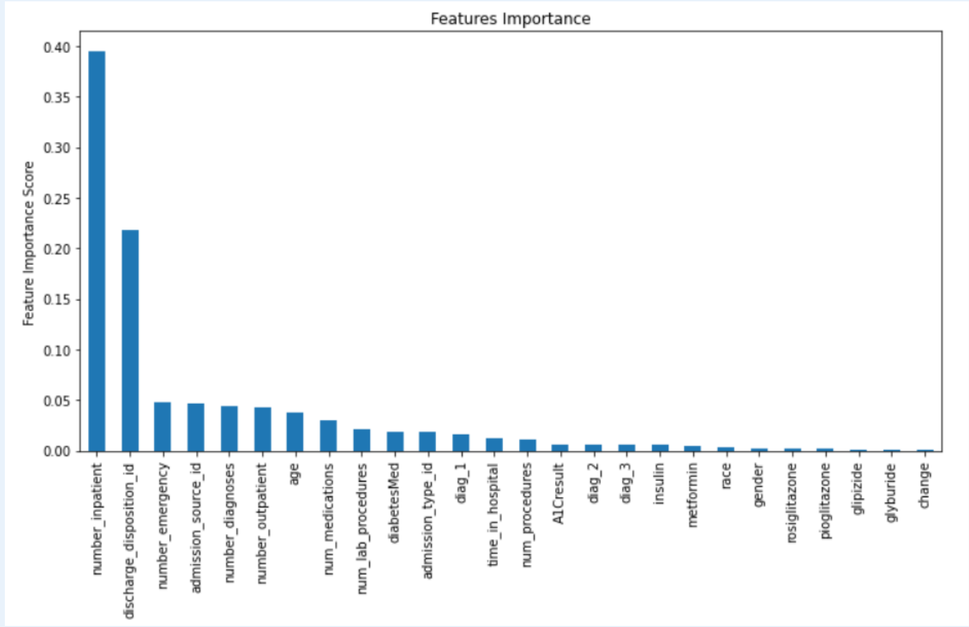
Machine Learning

Select the most important features

- We implement our models with 24 variables. We will choose the most important ones for the Gradient Boosting Model.

- We can notice that the time spent in hospital is not an important feature in the model. This surprised us and we did further research on this variable to understand. We can see that no matter if the patient will be readmitted or not, the distribution of the time spent in the hospital is the same.

- We finally drop features that we decided to exclude, and we scale and implement the Gradient Boosting model on the new data.



NOYELLE Amélie –KEENER Jeremy – PENIGUEL Josselin

Conclusion

- We chose this dataset because it is the one that interested us the most, the problematic posed related to the medical field was particularly topical.
- We are satisfied with a prediction accuracy of about 60% for a variable that may seem abstract. We were able to see that the dataset needed to be studied in greater depth. Some variables that seem to have an important weight in the prediction can turn out to be useless or even bias our model.
- This project taught us to use several machine learning models and their application in concrete contexts.
- We would like to thank Mr. SABRY for his guidance and involvement during this semester.