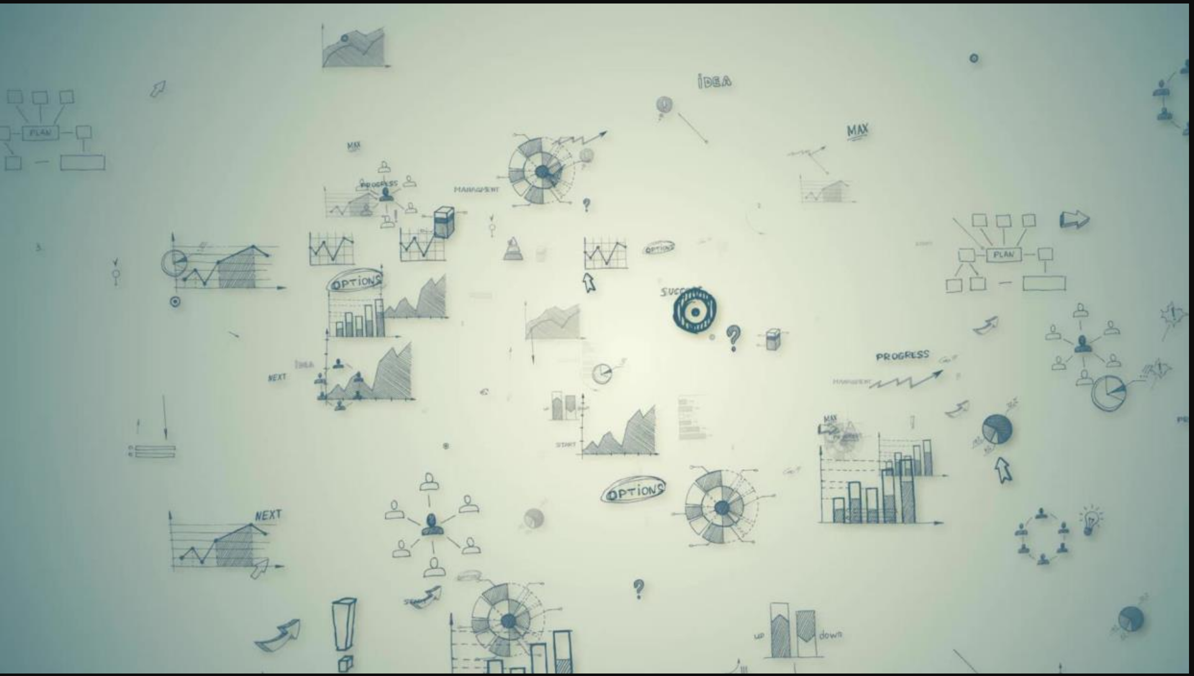
Python project : Diabete in the US

10 years (1999-2008) of clinical care at 130 US hospitals

From UCI Machine Learning repository

LE LORIER Lucas
NGALAMULUME Jonathan
MOHAMMAD Aima

DIA4



What you'll see

- 1. Dataset overview
- 2. Preprocessing strategy
- 3. Ideas on the topic
- 4. Settling for a Machine Learning problematic
- 5. A brief demonstration of our model on API

Pre-processing

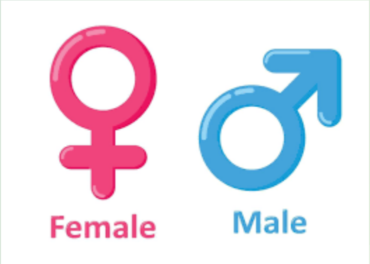
"Databases of clinical data contain valuable but heterogeneous and difficult data in terms of missing values" (p.2 description file)

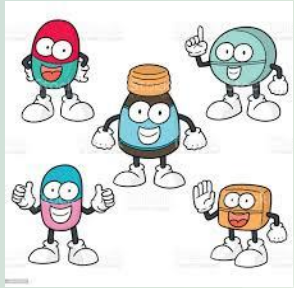
Diabetic_dataset's overview

TABLE 1: List of features and their descriptions in the initial dataset (the dataset is also available at the website of Data Mining and Biomedical Informatics Lab at VCU (http://www.cioslab.vcu.edu/)).

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: [0, 10), [10, 20), . . ., [90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: ">200," ">300," "normal," and "none" if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: ">8" if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: "change" and "no change"	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: "yes" and "no"	0%
24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed	0%

```
[ ]  df.shape

     (101766, 50)
```





- Nombre de procedure
- Changement de médicamebt
- Spécialité du médecin
- Type de diagnostique
- etc

Pre-processing

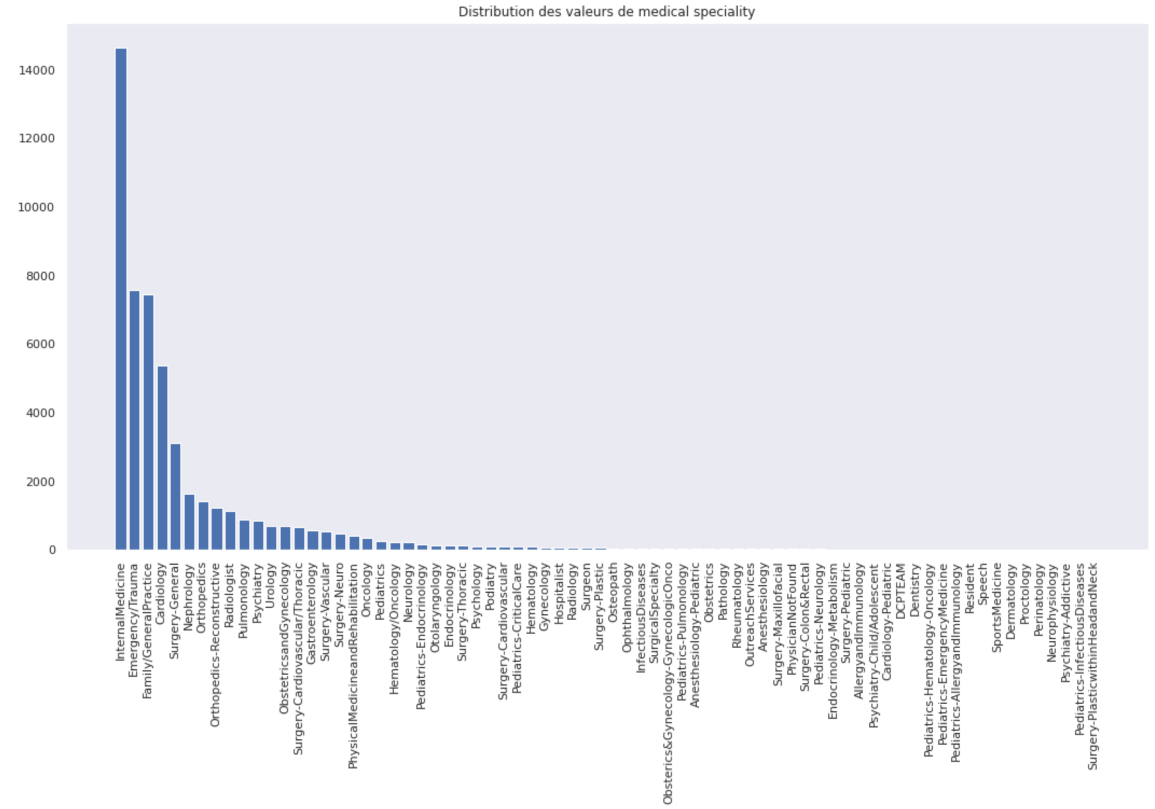
Processing of almost empty columns and poor data columns

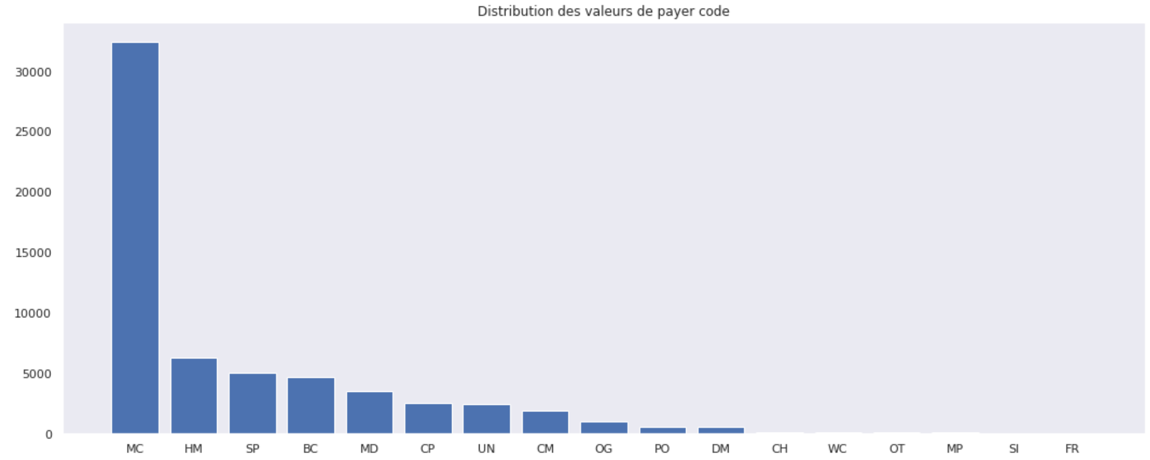
```
df_description[df_description["Suggestion"] == "Remove_columns"]
```

	Field	% Missing	Qty Unique	Suggestion
5	weight	0.9686	10	Remove_columns
10	payer_code	0.3956	18	Remove_columns
11	medical_specialty	0.4908	73	Remove_columns
39	examide	0.0000	1	Remove_columns
40	citoglipton	0.0000	1	Remove_columns

Pre-processing

medical_speciality and payer_code





Pre-processing

Processing of almost filled columns

```
df_drop_lines
```

	Field	% Missing	Qty Unique	Suggestion2
2	race	0.0223	6	Remove lines
18	diag_1	0.0002	717	Remove lines
19	diag_2	0.0035	749	Remove lines
20	diag_3	0.0140	790	Remove lines

Pre-processing

Race column

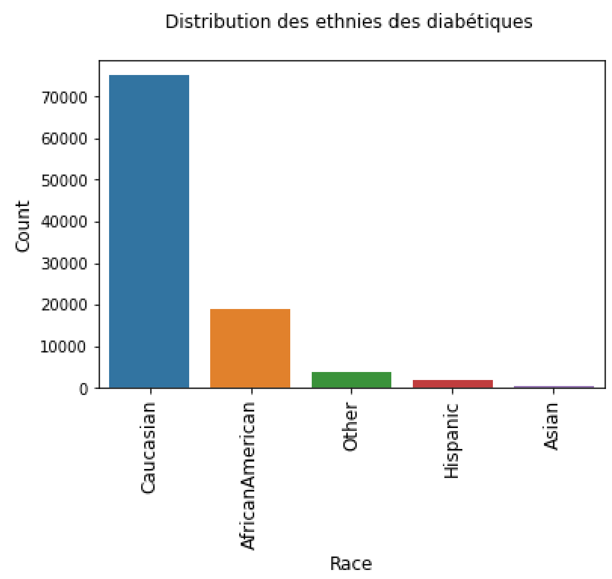
Fill NaN values in 'Other' race

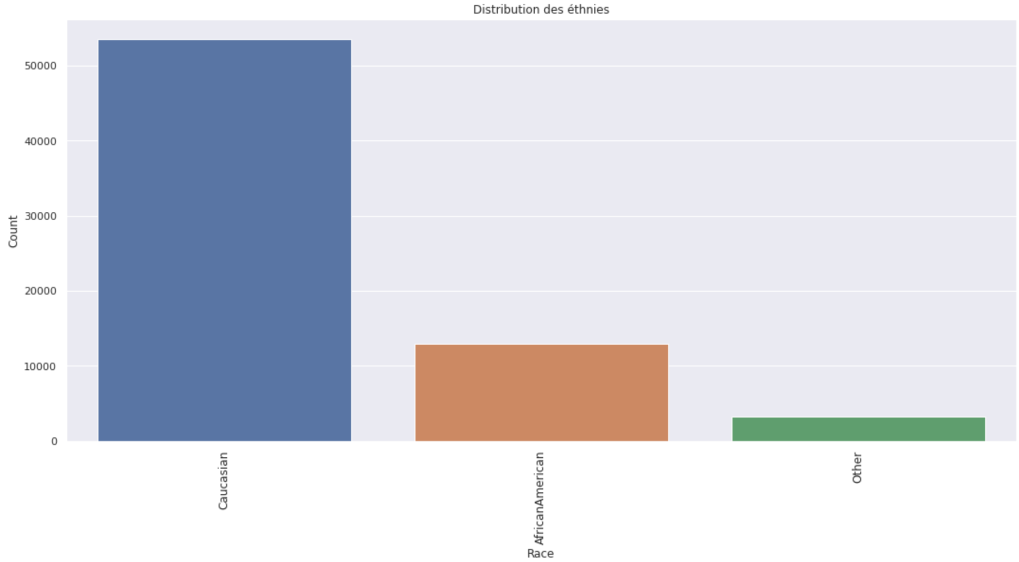
↓

Merge underepresented races in 'Other' race

↓

2 majors races : Caucasian ans AfricanAmerican

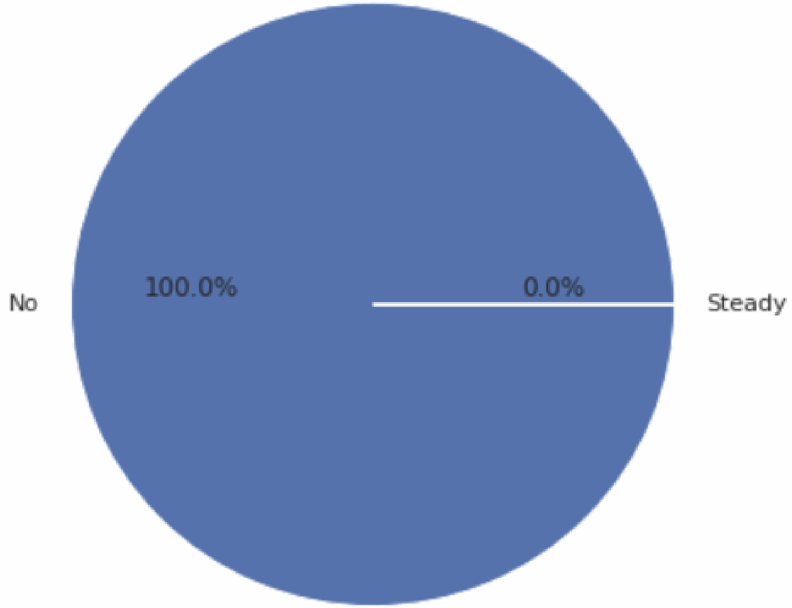




Pre-processing

- Medecine intake columns

```
No          71513
Steady          2
Name: metformin-rosiglitazone, dtype: int64
```



Pre-processing

Dropping non interesting columns/ with nearly redundant information

```
df=df.drop(columns=['discharge_disposition_id','admission_source_id','admission_type_id'])

df
```

- Discharge disposition: Nominal Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available

- Admission source : Nominal Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital

- Number of lab procedures : Numeric Number of lab tests performed during the encounter

- Number of procedures Numeric Number of procedures (other than lab tests) performed during the encounter

Pre-processing

Creation number of encounters column /drop id encounters column/ set index as patients id columns

```
a=df.groupby('patient_nbr')['encounter_id'].count()
```

```
df=df.sort_values(by=['patient_nbr'])
df=df.drop_duplicates(subset=['patient_nbr'], keep='last')
df['nb_encounter']=list(a.values)
df=df.drop(['encounter_id'], axis=1)
df
```

Merge columns nb lab procedures & nb procedures / Drop columns nb lab procedures & nb procedures

```
## Fusion des colonnes procedures :

df['nb_procedures'] = df.loc[: , ['num_procedures', 'num_lab_procedures']].sum(axis=1)
df=df.drop(columns=['num_procedures', 'num_lab_procedures'], axis=1)
df
```

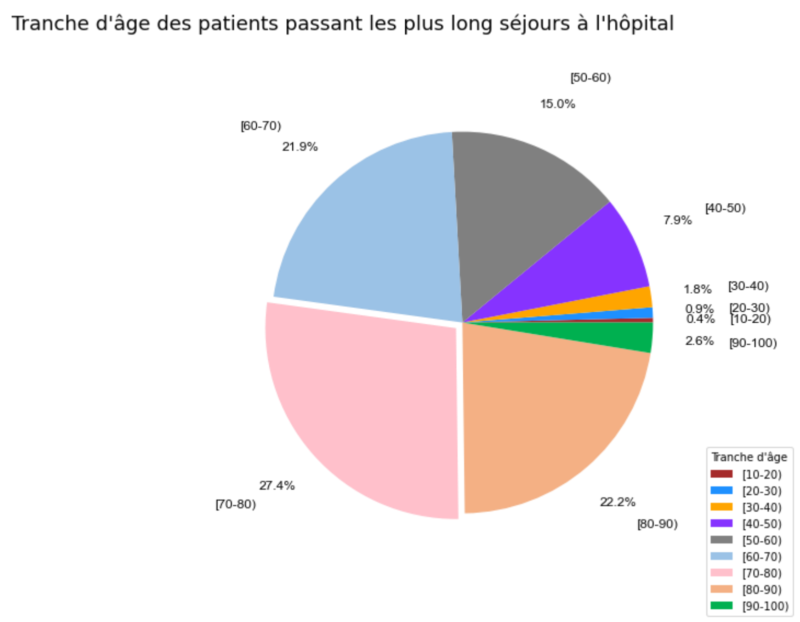
Merge columns umber outpatient number inpatients / Drop columns umber outpatient number inpatients

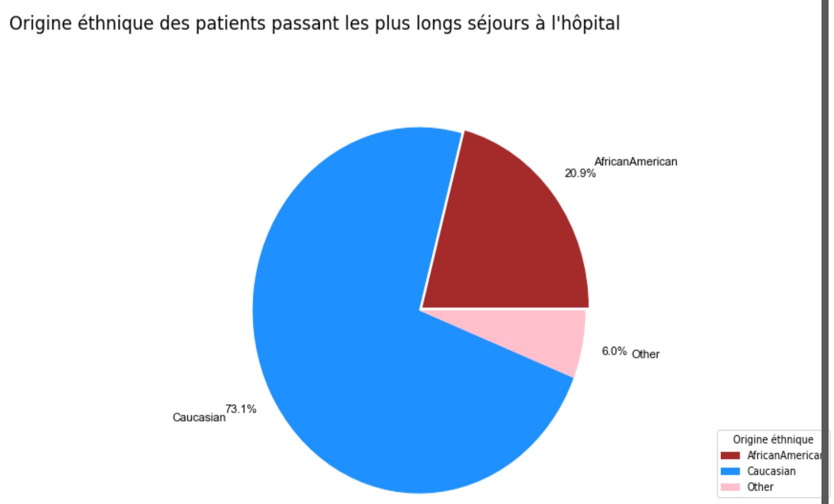
```
## Fusion des colonnes patient :

df['nb_patients'] = df.loc[: , ['number_outpatient', 'number_inpatient']].sum(axis=1)
df=df.drop(columns=['number_outpatient', 'number_inpatient'], axis=1)
df
```

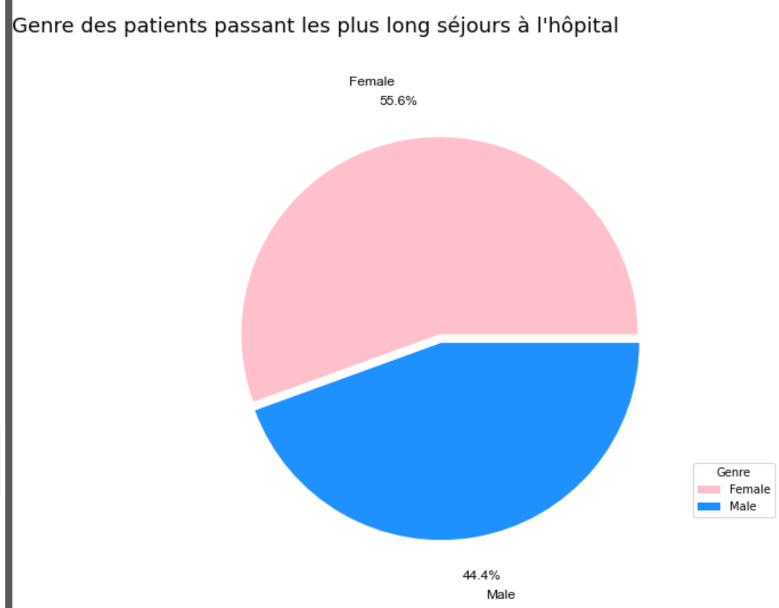
Pre-processing

Binarisation of our output column readmission :

```
# on définit valeurs oui non readmission:
df['readmitted'] = df['readmitted'].replace({"<30":'YES',
                                            ">30":'YES'})
```

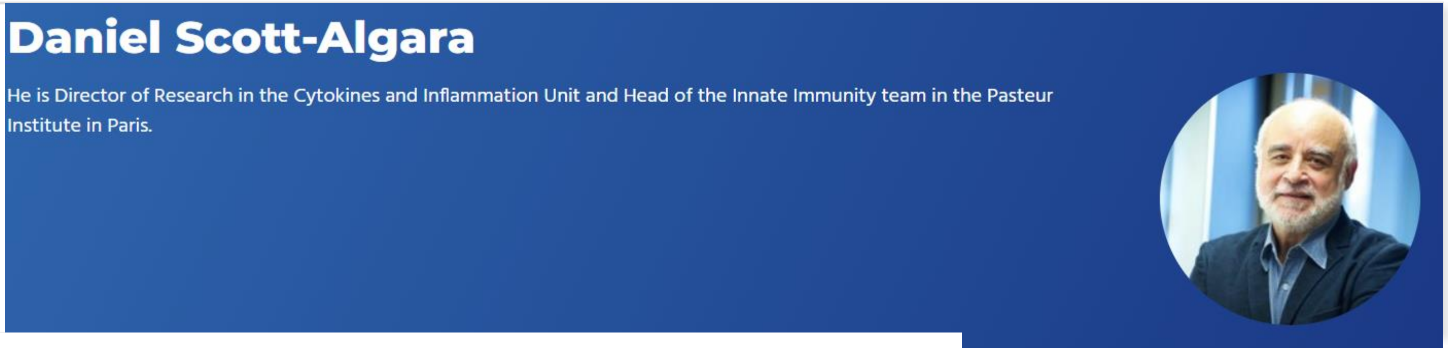




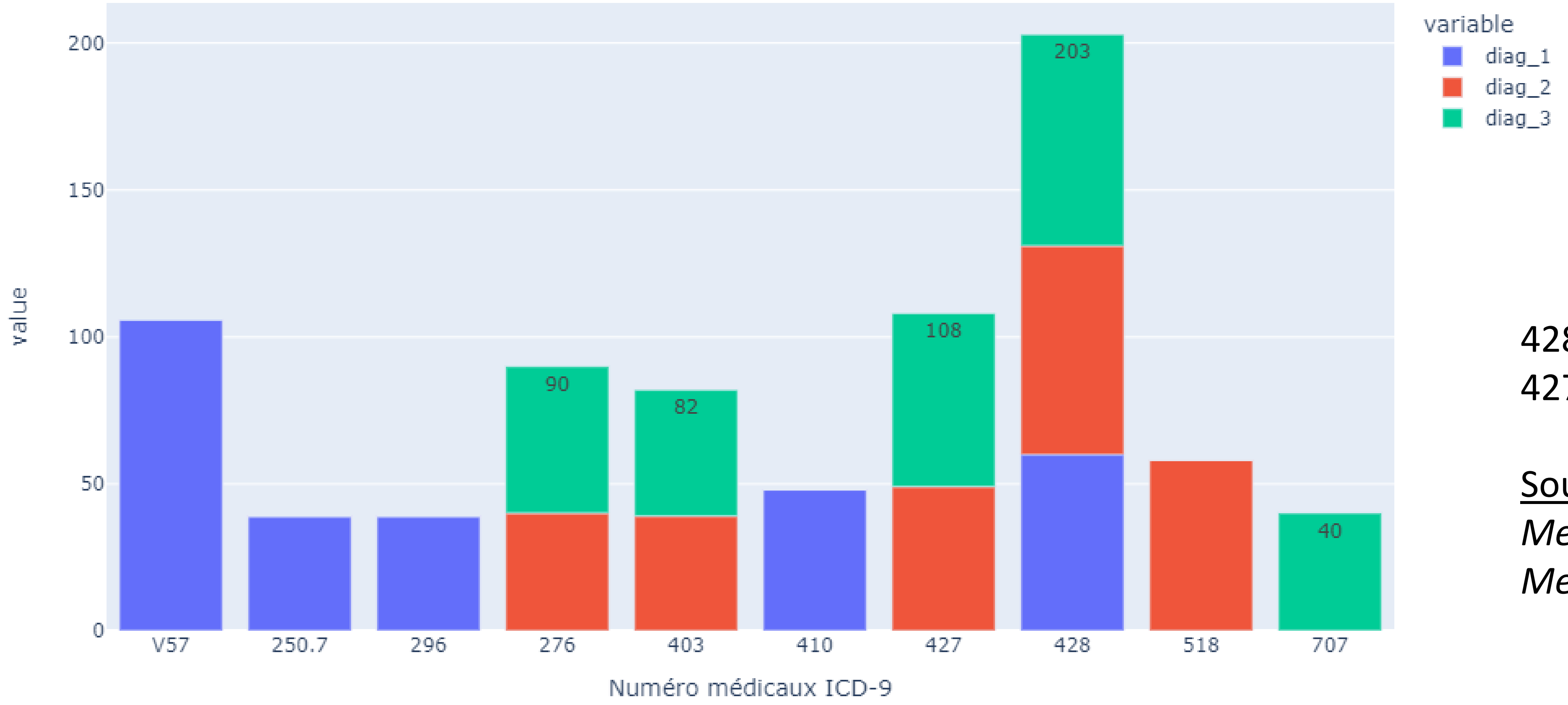


Analysis : what is the profile of the patient having the longest hospital stay ?

Analysis 1 : what is the profile of the patient having the longest hospital stay ?

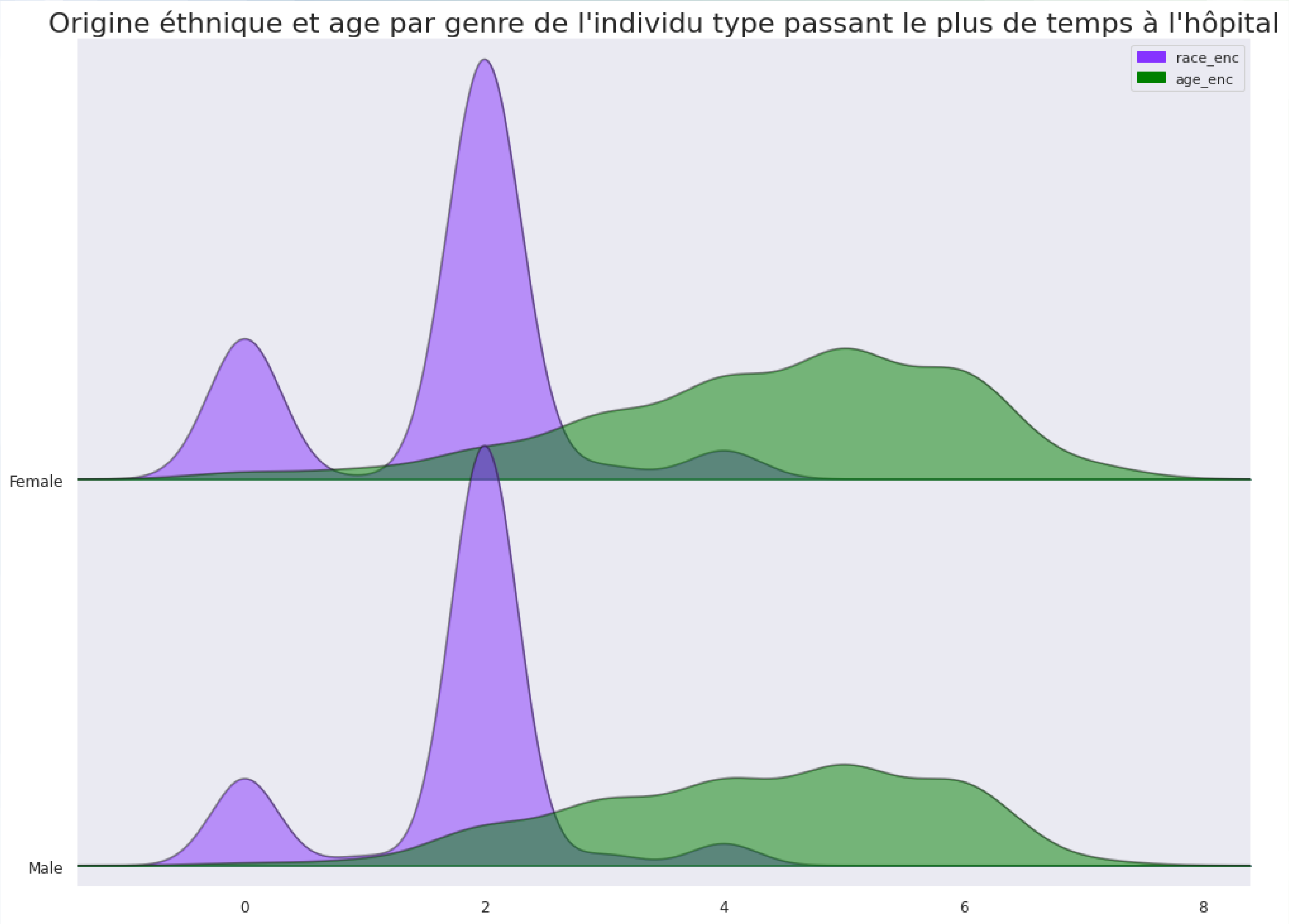


ICD-9 !

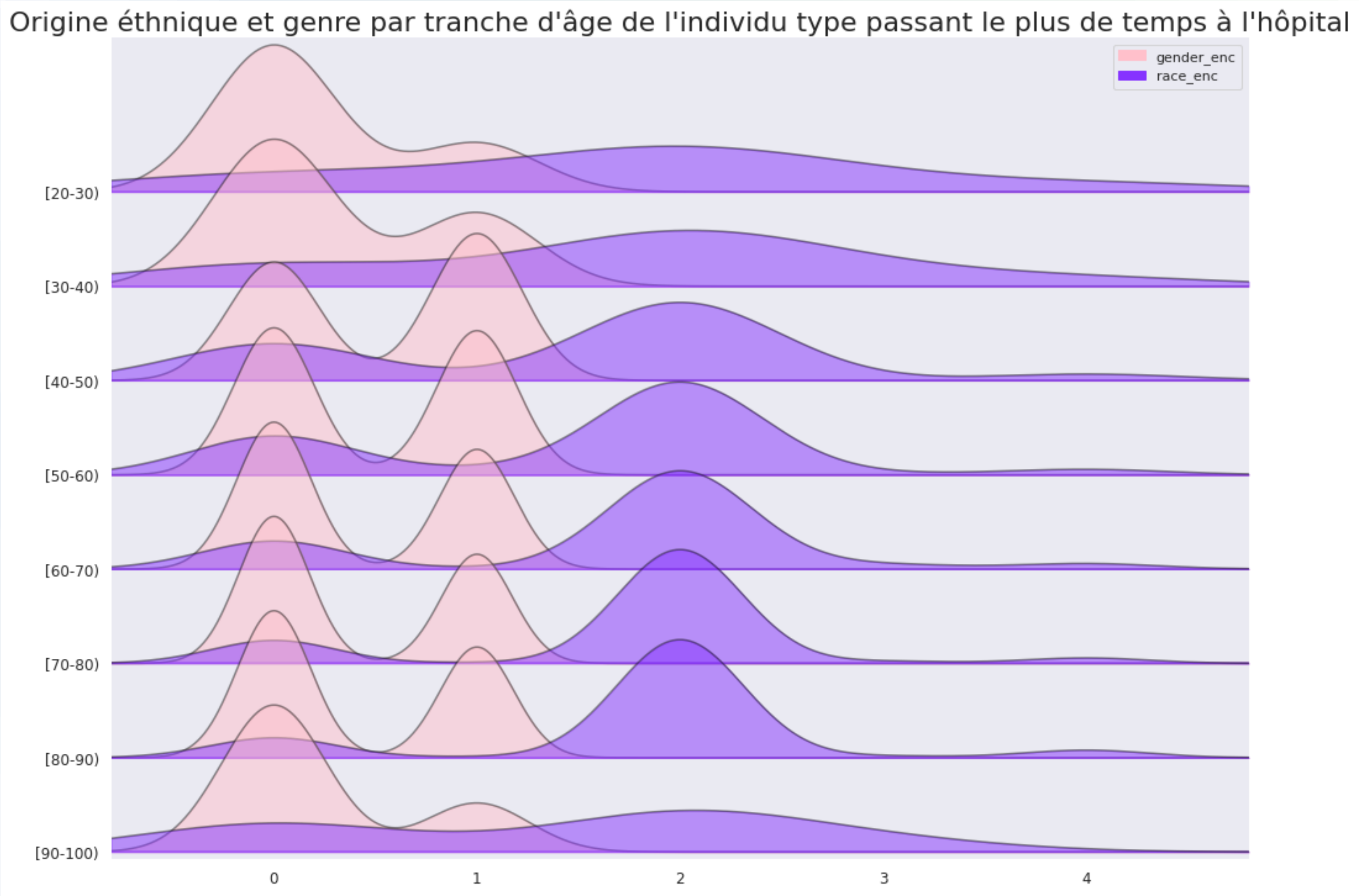


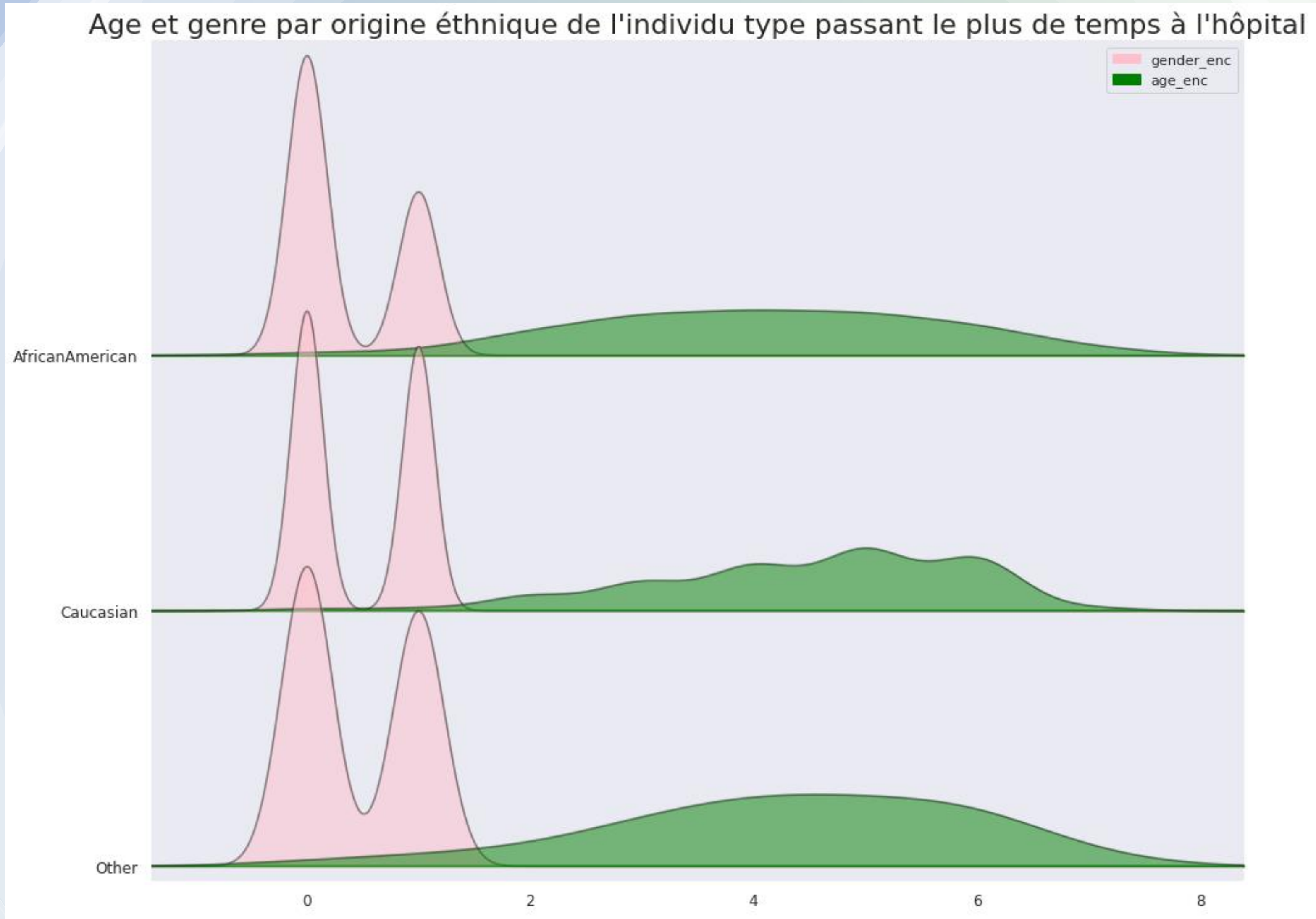
428 : heart failure
427 : Cardiac dysrhythmias

Source: *health.gov> Data Methodology and Measurement*



Caucasian=2. African-American=0. Female=0. Male=1


Origine éthnique et genre par tranche d'âge de l'individu type passant le plus de temps à l'hôpital
gender_enc
race_enc
[20-30)
[30-40)
[40-50)
[50-60)
[60-70)
[70-80)
[80-90)
[90-100)
0
1
2
3
4



Age et genre par origine éthnique de l'individu type passant le plus de temps à l'hôpital
gender_enc
age_enc
AfricanAmerican
Caucasian
Other
0
2
4
6
8

Conclusion : the profil of the typical US diabetic patient

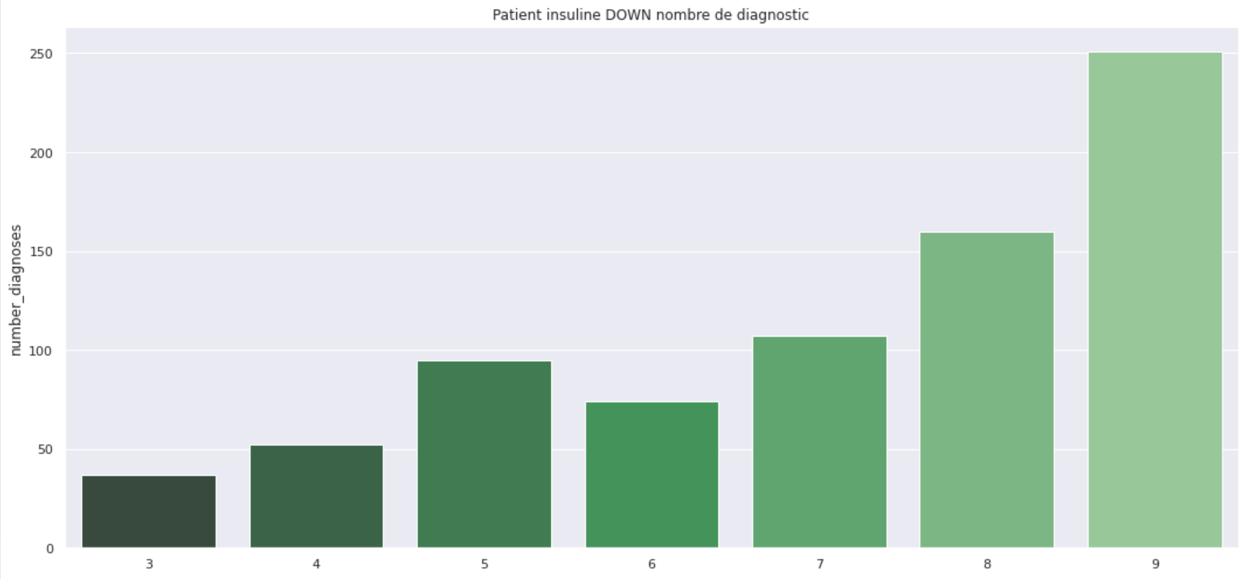
Feedbacks from medical experts : our journey to understand the data and set for a Machine Learning problem

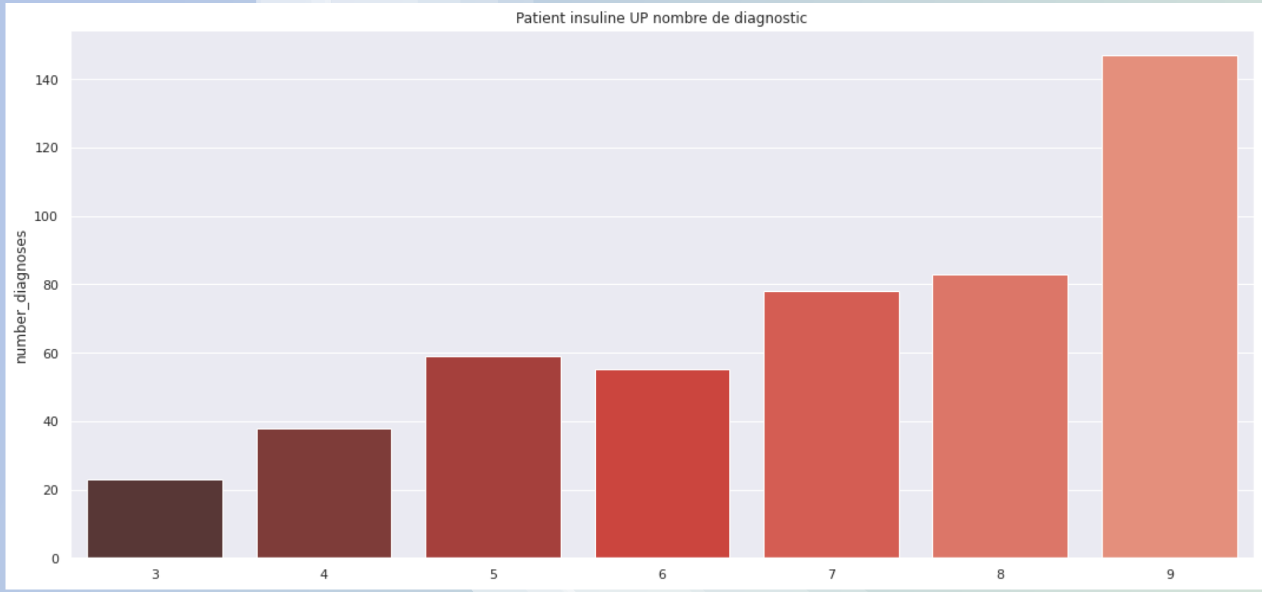
- Diabetic II : The patient takes only oral medication. But if we predict insulin for this patient before the age of 30, he becomes a type 1 diabetic. So, it is possible that a patient can change type of diabete
- Type 1 diabetes has only one treatment: insulin replacement. There is no other treatment for Type 1 diabetes and without insulin, death is very likely. Whereas Type 2 diabetes can be managed with diet, weight loss, medications, and/or insulin.
- The more the patient is having diagnosis, the better will go his/her diabete situation.
- Its interessing to study the prediction of patient readmission.

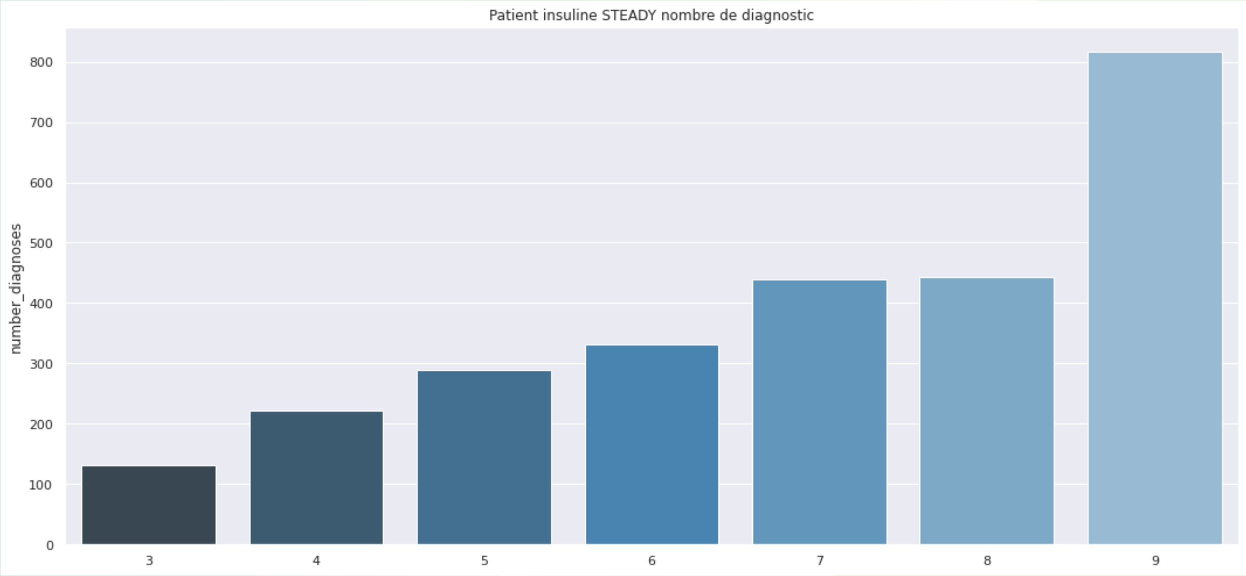
Daniel Scott-Algara

He is Director of Research in the Cytokines and Inflammation Unit and Head of the Innate Immunity team in the Pasteur Institute in Paris.

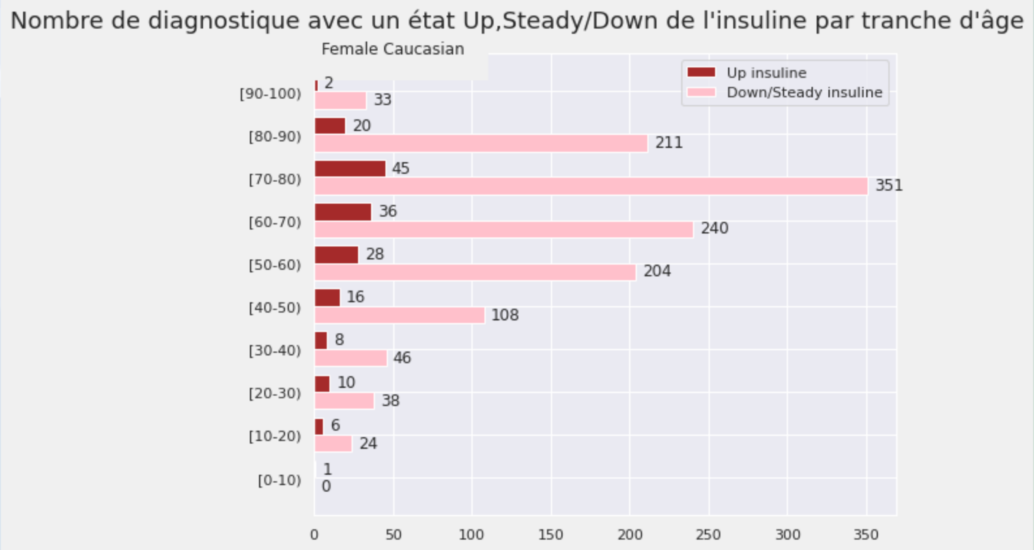
Analysis 2: *"The more the patient is having diagnosis, the better will go his/her diabete situation."*

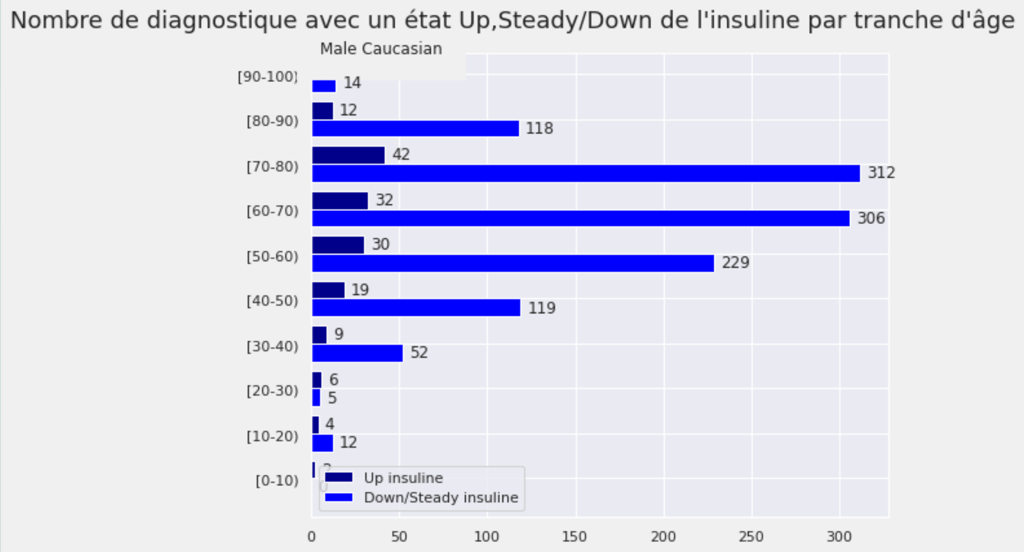


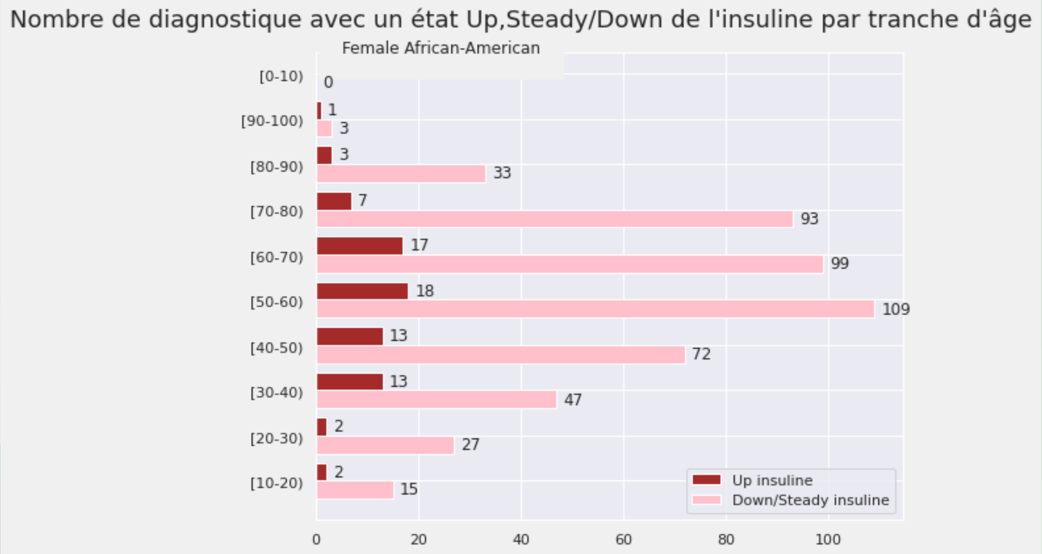


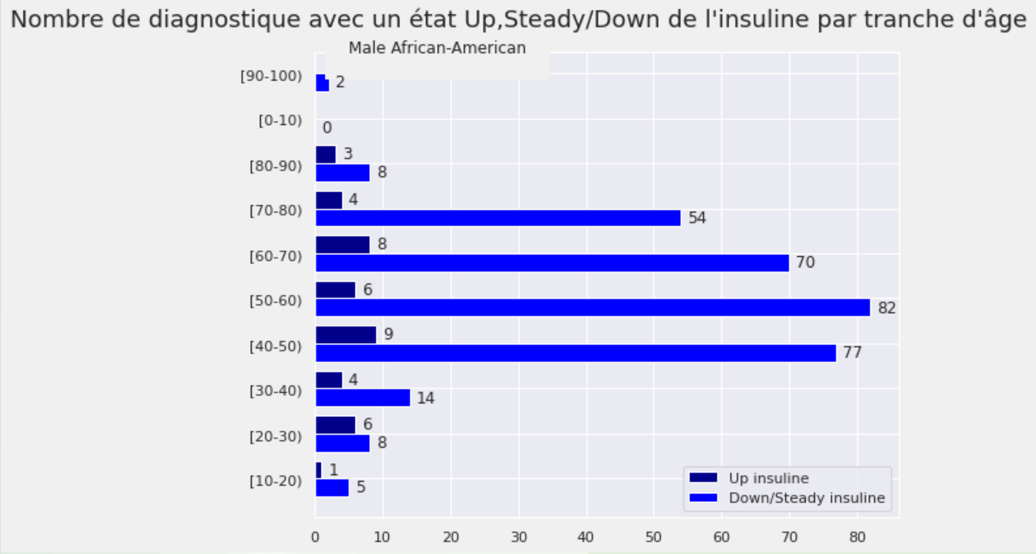


Analysis 2: *"The more the patient is having diagnosis, the better will go his/her diabete situation."*



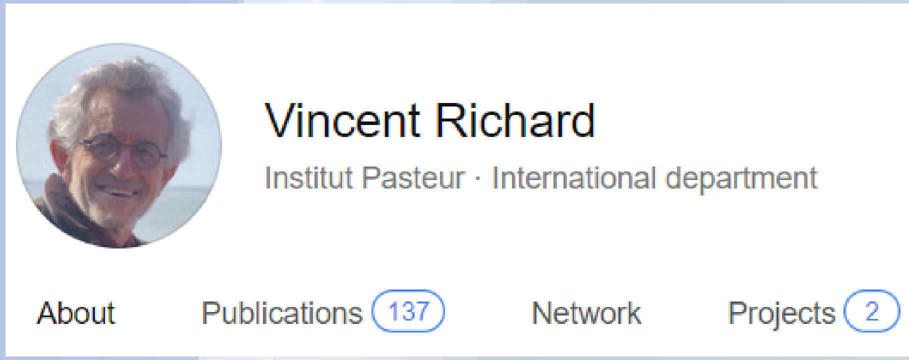


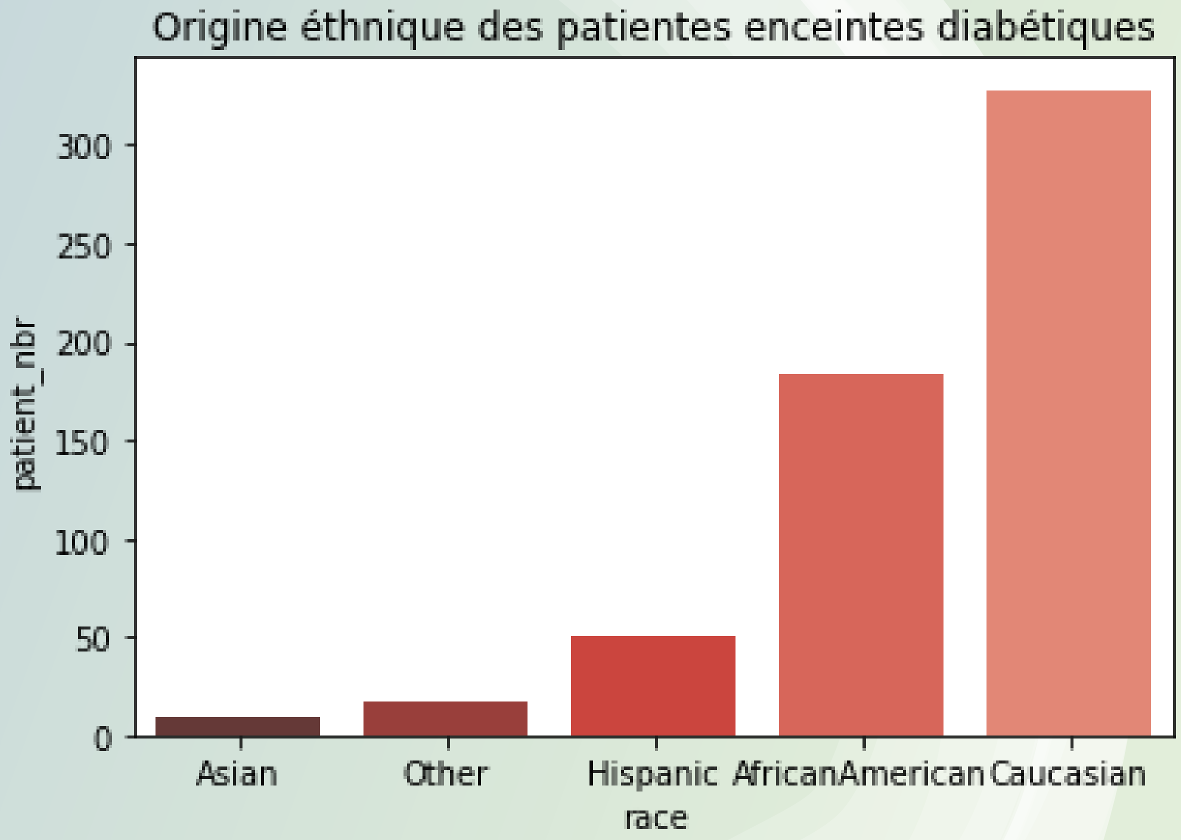




Analysis 3 : Diabete and pregnancy

- "Diabete appears frequently during pregnancy and is part of the complications that could lead to the readmission of the mother-to-be."
- ICD-9 codes from 630 to 679 refer to complications during pregnancy.






```
fem=df[df.gender=='Female']
```

```
femp=fem[('630'<fem.diag_1) & (fem.diag_1<'679')|('630'<fem.diag_2) & (fem.diag_2<'679')|('630'<fem.diag_3) & (fem.diag_3<'679')
femp.shape
```

```
(651, 45)
```

After reflection we chose to predict the readmission or not of a patient

Model results

	model	accuracy	best_params
0	logistic_regression	0.744381	{'C': 10}

```
AUC:0.654
Accuracy:0.744
Recall:0.381
Precision:0.726
confusion_matrix:
[[10846   848]
 [ 3661  2249]]
```

Demo API

Thank you !