Estimation of obesity levels based on eating habits and physical condition

Final project in python for data analysis

Louis COUSIN
Fanny DEBORD

Introduction

The objective of this project is to create a model capable of determining a person’s level of obesity based on their personal information. This model will be made available on API.

The different levels of obesity are : insufficient weight, normal weight, overweight level I, overweight level II, obesity type I, obesity type II, and obesity type III

Research Resources – The data set

The dataset is available here : https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+#

And its description is available here ; sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub

The dataset contains 16 variables. 15 features:

- Gender
- Age
- Height
- Weight
- Family History with Overweight
- Attributes related with eating habits (6)
- Attributes related with the physical condition (4)

One target:

- NObeyesdad, equal to NObesity

Data set summary – data dictionnary (features)

Columns	Related question	Value
Gender	What is your gender ?	{« Female », « Male »}
Age	What is your age ?	Int
Height	What is your height?	Float in meters
Weight	What is your weight ?	Float in kilograms
Family_history_with_overweight	Has a family member suffered or suffers from overweight?	{« yes », « no »}
FAVC	Do you eat high caloric food frequently?	{« yes », « no »}
FCVC	Do you usually eat vegetables in your meals?	{« Never », « Sometimes », « Always »}
NCP	How many main meals do you have daily?	int
CAEC	Do you eat any food between meals?	{« No »,« Sometimes », « Frequently », « Always »}
SMOCKE	Do you smoke ?	{« Yes », « no »}

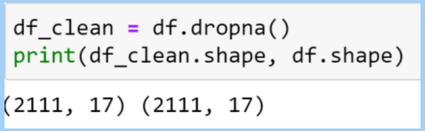
Columns	Related question	Value
CH2O	How much water do you drink daily?	Float in liter
SCC	Do you monitor the calories you eat daily?	{« Yes », « No »}
FAF	How often do you have physical activity?	Float (per week)
TUE	How much time do you use technological devices?	Float (hours)
CALC	how often do you drink alcohol?	{“no”,“Sometimes”, “Frequently” ,”Always“}
MTRANS	Which transportation do you usually use?	{“Automobile”, “Motorbike”, “Bike”,,“Public,Transportation’,Walking”}

Data analysis

In this part we will see the data preprocessing for graphic analyzes :

- Data cleaning : part 1
- Data analysis :

Data cleaning : part 1



```
df_clean = df.dropna()
print(df_clean.shape, df.shape)
```

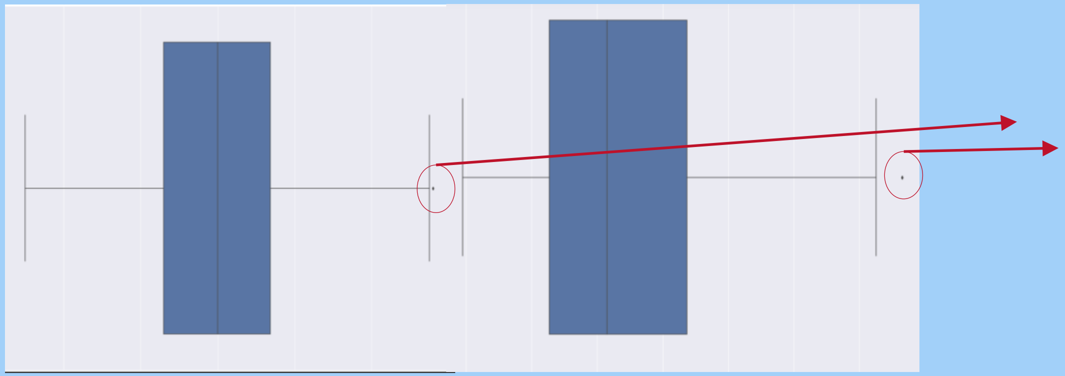
```
(2111, 17) (2111, 17)
```

We can see that they have the same shape, so there is no missing value: we keep **df** for our analyse.

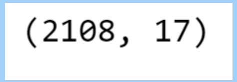
For non-categorical variables, we need to check if there is any outliers. There 3 concerned variables are Age, Height and Weight.
Age is not a problem, because the study has been done with subjects between 14 and 61 years.

Height :

Weight :



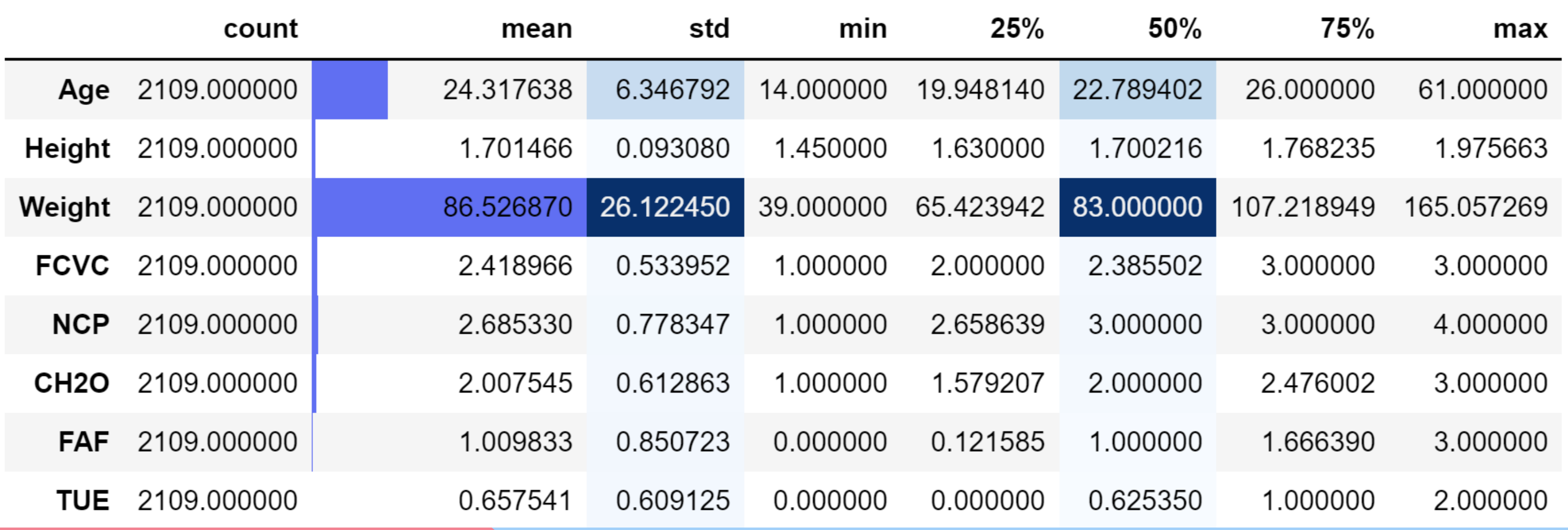
As you can see there is some outliers. We have decided to remove them in case it might distort our analyzes.



```
(2108, 17)
```

We delete 3 rows and now we have 2108 rows.

General view of quantitative data :

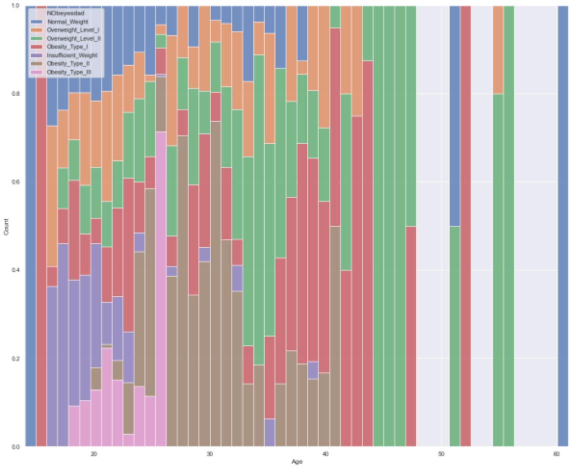


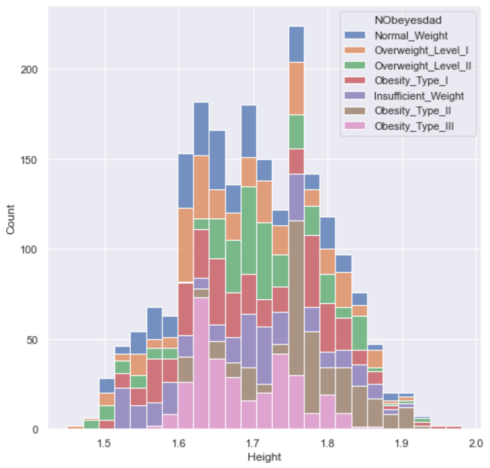
	count	mean	std	min	25%	50%	75%	max
Age	2109.000000	24.317638	6.346792	14.000000	19.948140	22.789402	26.000000	61.000000
Height	2109.000000	1.701466	0.093080	1.450000	1.630000	1.700216	1.768235	1.975663
Weight	2109.000000	86.526870	26.122450	39.000000	65.423942	83.000000	107.218949	165.057269
FCVC	2109.000000	2.418966	0.533952	1.000000	2.000000	2.385502	3.000000	3.000000
NCP	2109.000000	2.685330	0.778347	1.000000	2.658639	3.000000	3.000000	4.000000
CH2O	2109.000000	2.007545	0.612863	1.000000	1.579207	2.000000	2.476002	3.000000
FAF	2109.000000	1.009833	0.850723	0.000000	0.121585	1.000000	1.666390	3.000000
TUE	2109.000000	0.657541	0.609125	0.000000	0.000000	0.625350	1.000000	2.000000

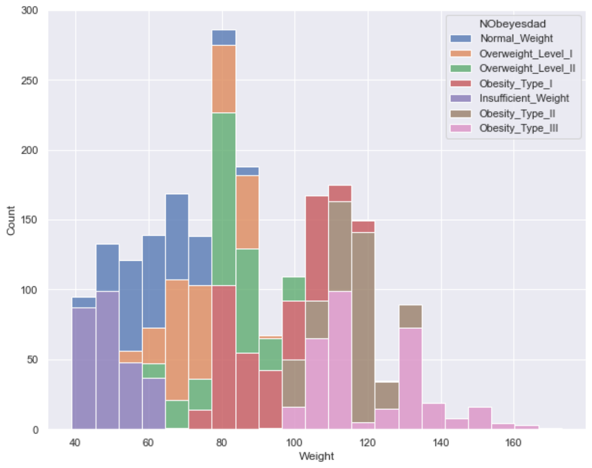
As we can see :

- The weight range is very wide and relatively well distributed.
- The people surveyed are relatively young
- The distribution of heights, weight and age let us think that the data set is representative of the population of Peru, Colombia, and Mexico (where the data come from) where the median age is 27 years

Variables related to levels of obesity



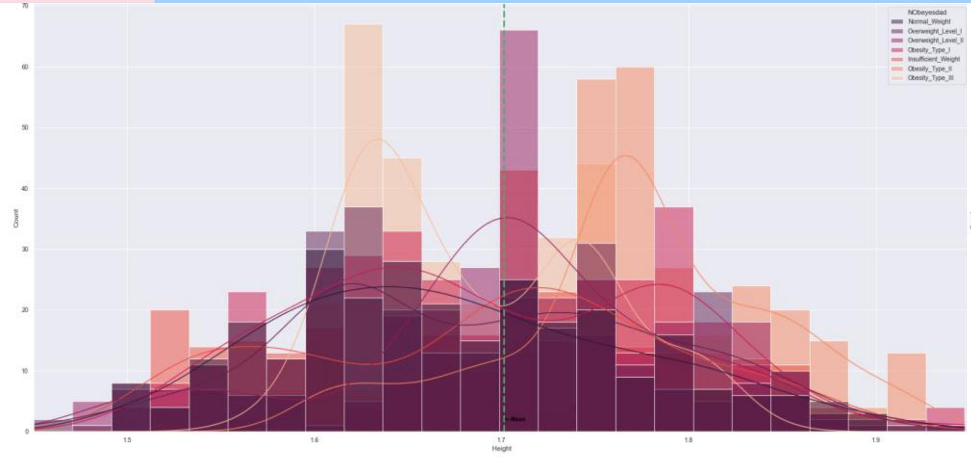


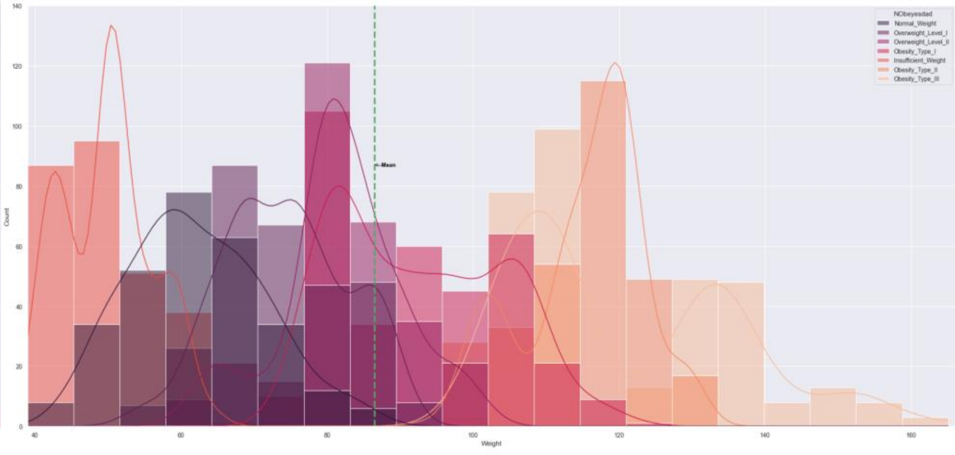


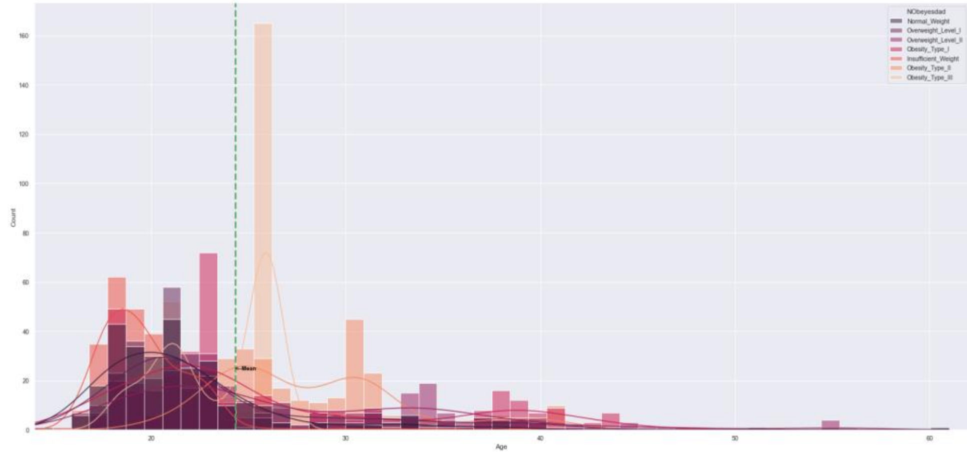
As we can see, Weight has a huge impact on the classification: an individual under 75kg cannot be considered as obese, which makes sense, but it considerably reduce the possibility choice. For the feature age, we can see that Obesity type 3 person have a life expectancy way less high than other: no case of type 3 are register over 30 years old.

All theses three physical criteria have an import impact on the repartition of the obesity level (It will be confirmed thanks to the correlation matrix)

Variables related to levels of obesity



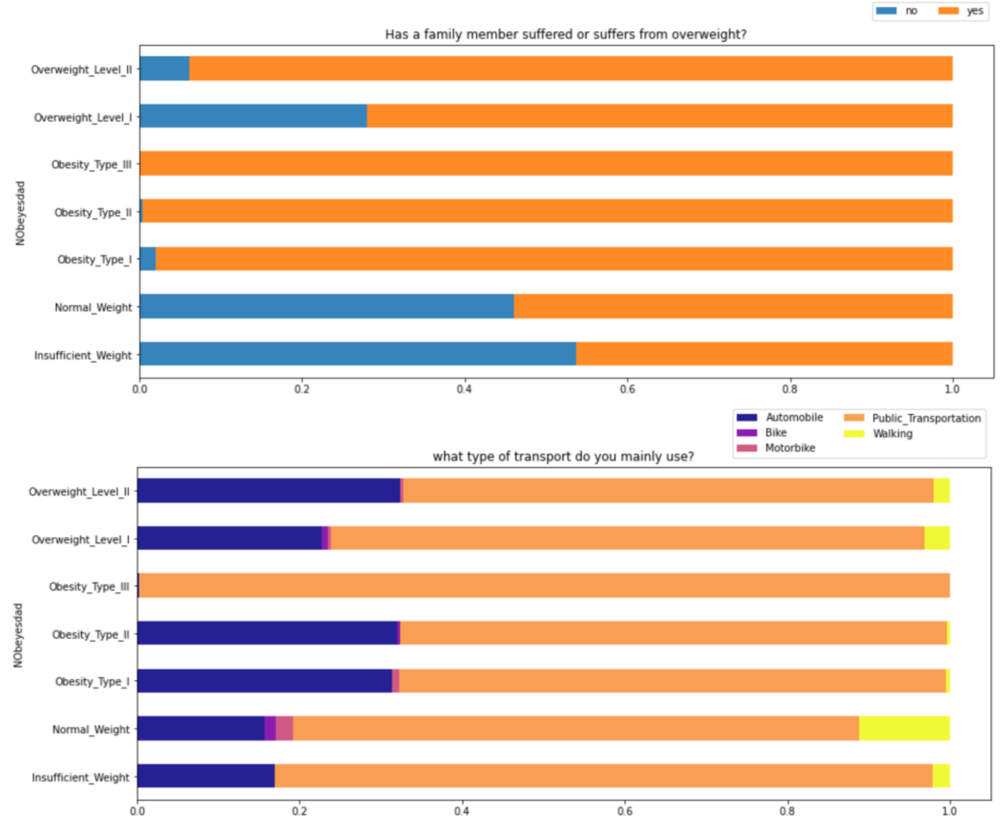




All distributions of obesity levels could be approached by one or more Gaussians each time. Note that for low non-normal levels (insufficient and above overweight) we will have to use several gausians. The only who is really close to a simple gausians is the weight of normal people.

This means that people with weight problems are mostly on "steps" while people without weight problems are evenly distributed over a certain range.

Variables related to levels of obesity

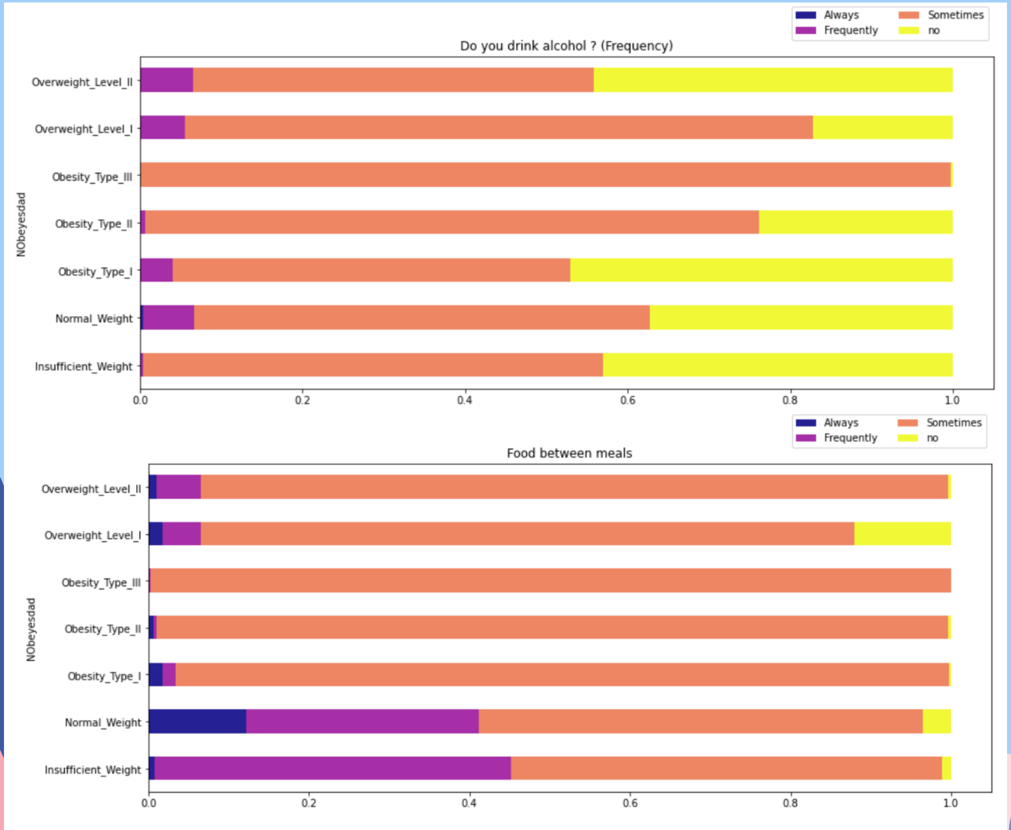


According to these graphics :

- on average, 92.6% of people who are obese or overweight have a loved one with overweight problems

- Overweight people use their cars more than people without weight problems

Variables related to levels of obesity



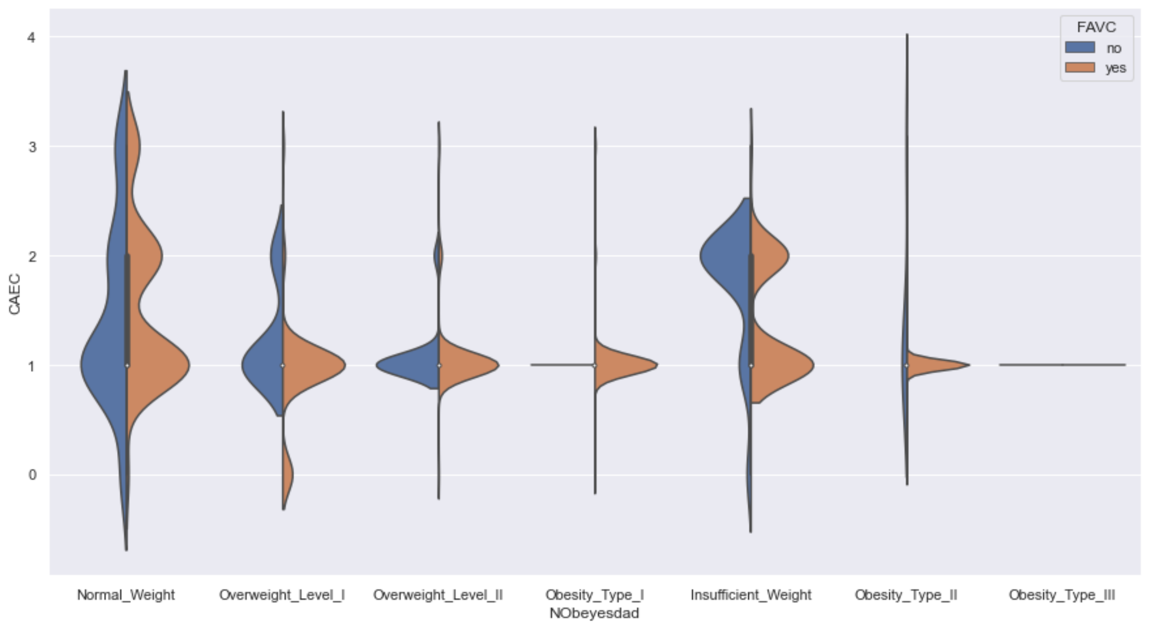
According to this graphics :

- It seems that people with weight problems drink alcohol. We will find more people who do not drink alcohol at all in people with underweight or normal weight

The link between alcoholism and obesity levels is relatively hard to investigate

- People who are underweight and normal tend to snack more between meals, but this may be due to the fact that the meals of those people there are less or not enough.

Variables related to levels of obesity

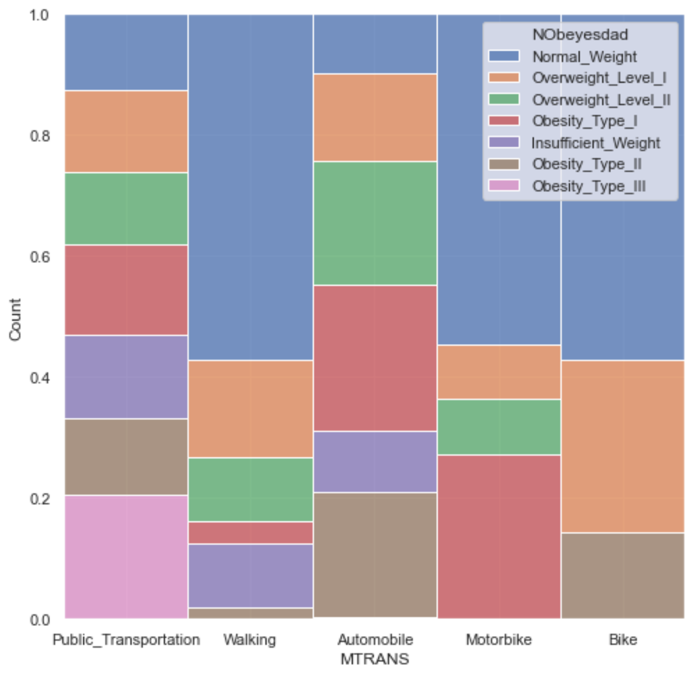


- Clearly show that if an individual eats high caloric food, it has more chance to be obese.
- As previoulsy seen, people who are underweight and normal tend to snack more between meals.

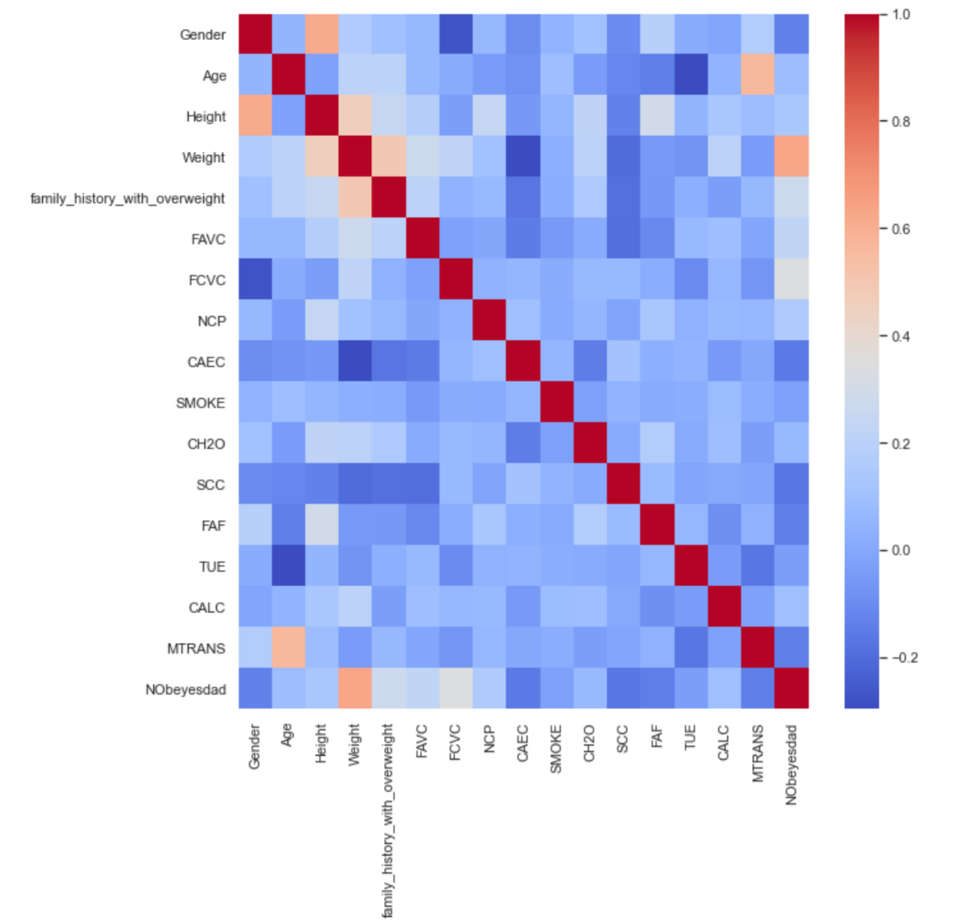
CAEC: Food between meals
FAVC: High caloric food

Variables related to levels of obesity

- The more nomad an individual is, the less it has chance to be obese.
- Most of people who use bike, motorbike or their feet are considered as normal wheight person, which makes sens. The burn more calories thought the day.



Variables related to levels of obesity



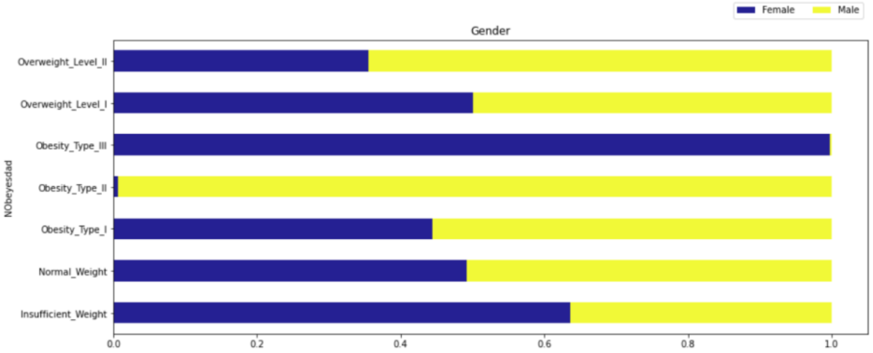
According to this graphics, the most influential parameters on a person's level of obesity are :

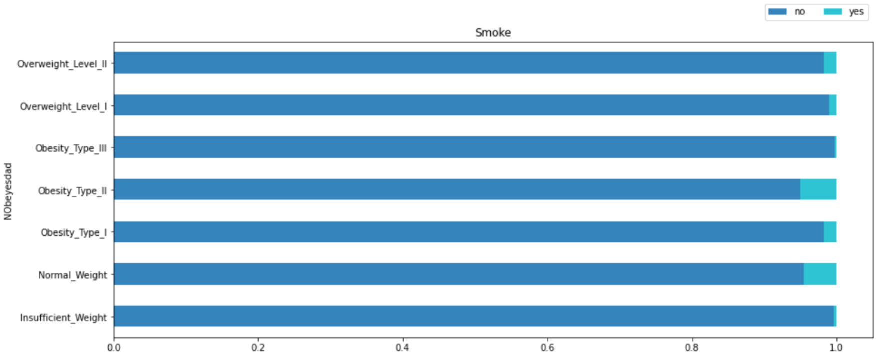
- ➢ Their weight
- ➢ Their height
- ➢ How often they eat vegetables
- ➢ Their genetics (family with overweight problems)
- ➢ How often they eat high-calorie foods
- ➢ Their nomber of main meals
- ➢ How often they drink alcohol

However, the variables are relatively poorly correlated (less than 0.4 correlation) with the level of obesity (except for the weight).

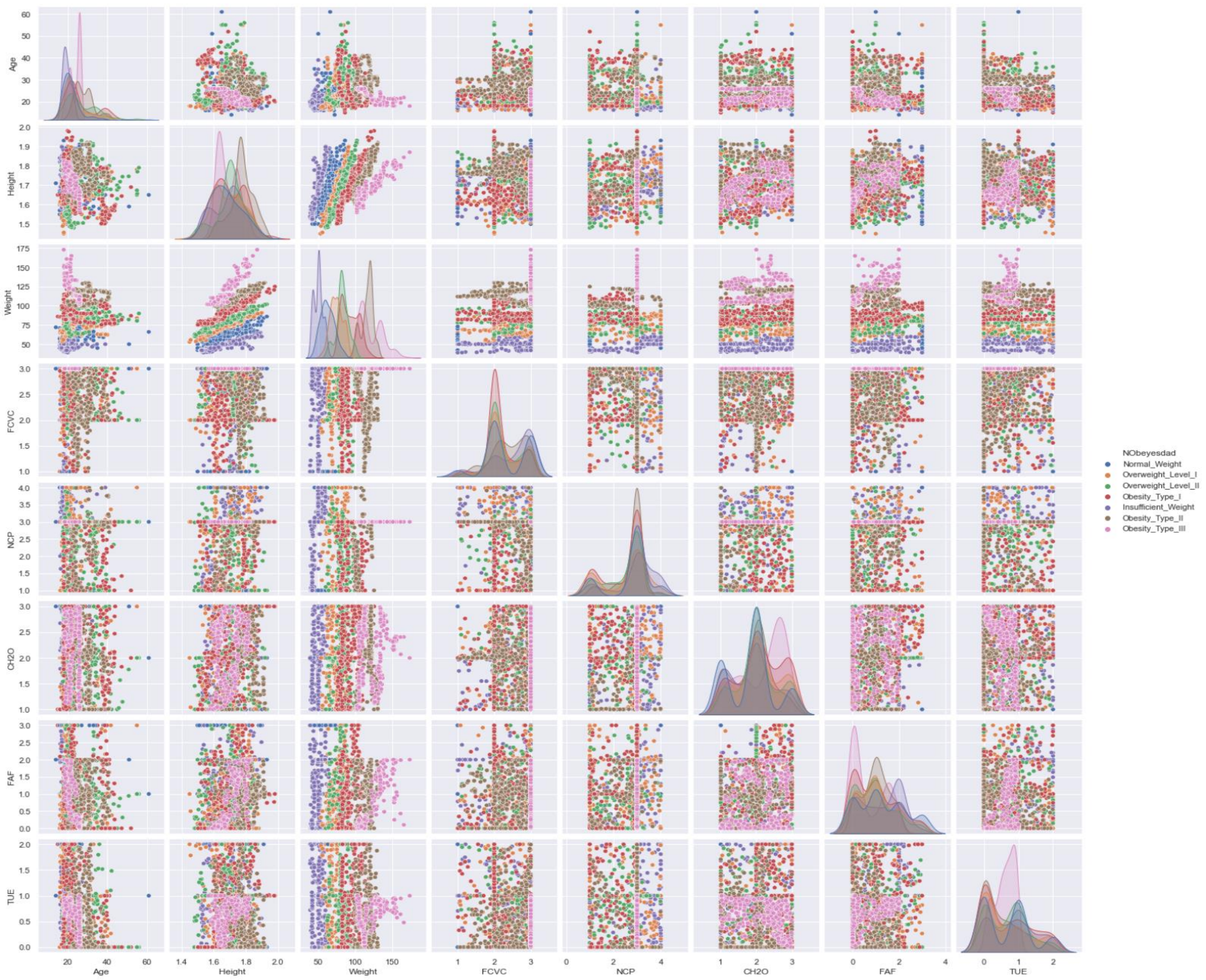
Furthermore, We can see that every variable is globaly lineraly independant from each other. Because of that, ACP will not be efficient, so we do not perform it. We will select the variable correlated with less than 30%.

Variables not related to levels of obesity

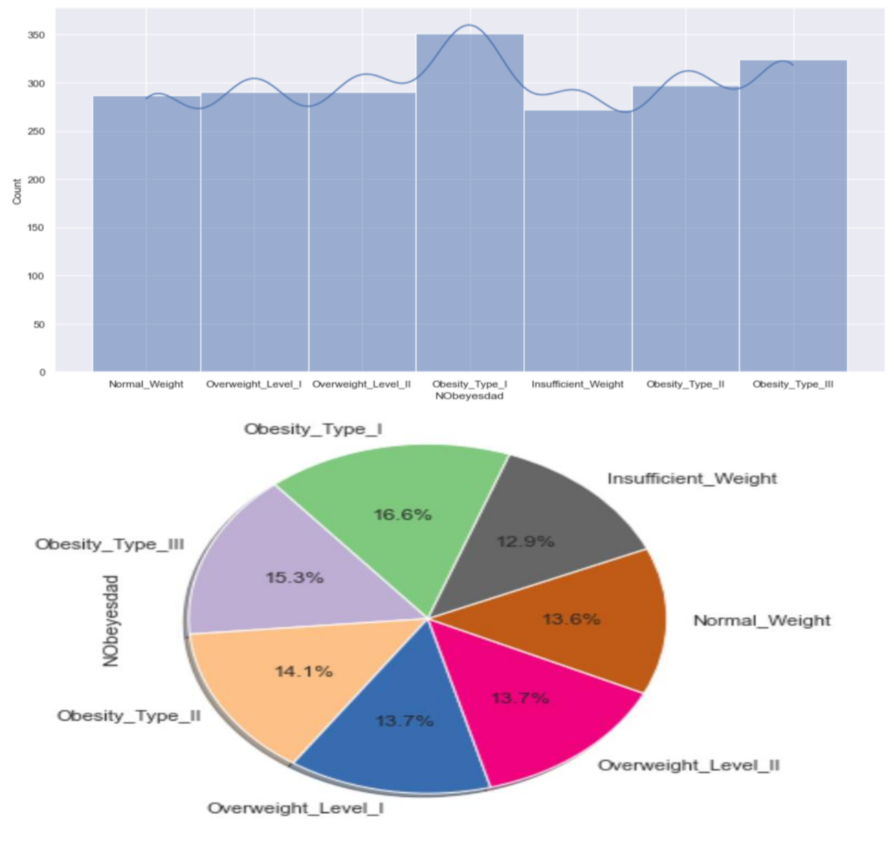




General analysis on variable :



Target analysis :

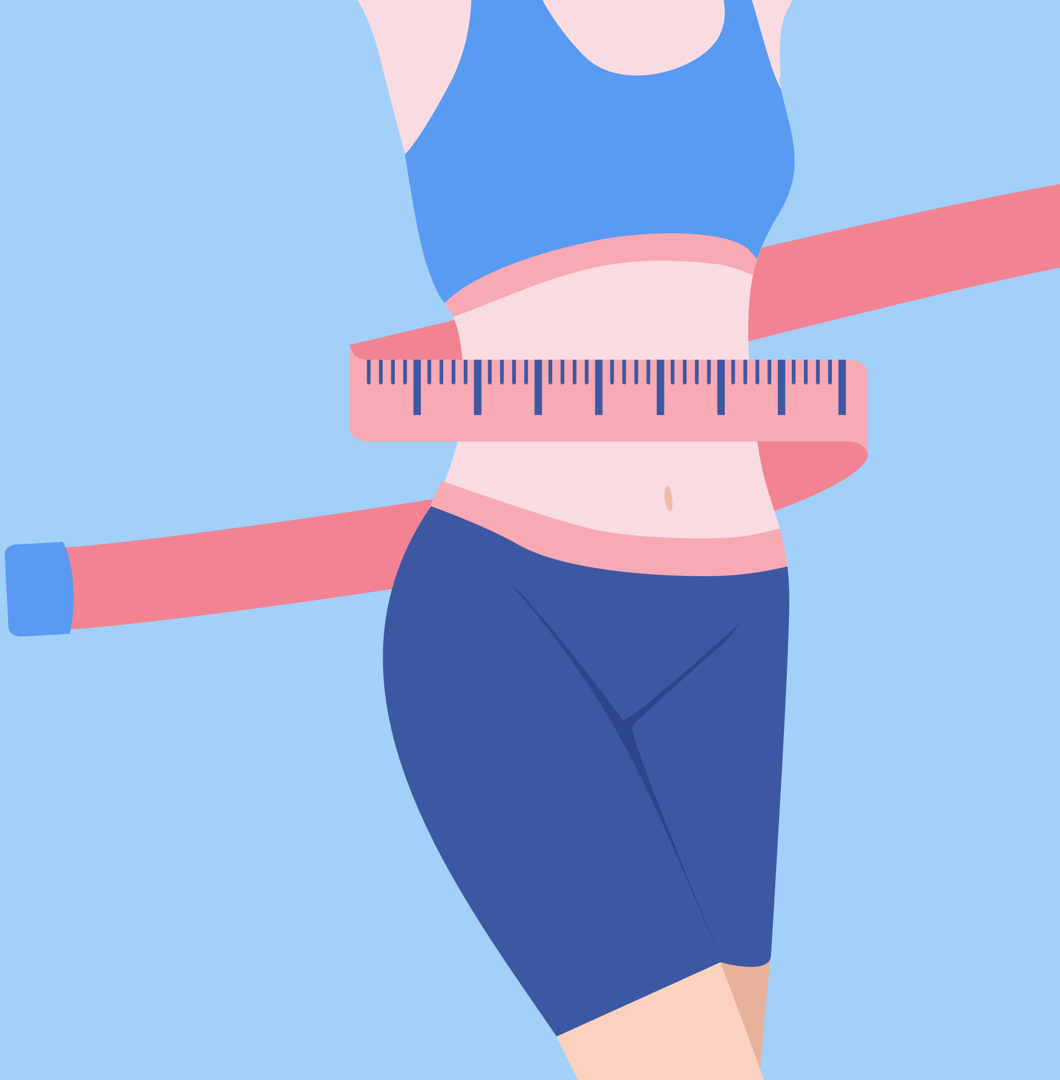


The target is well distributed across the different classes.

Modelization

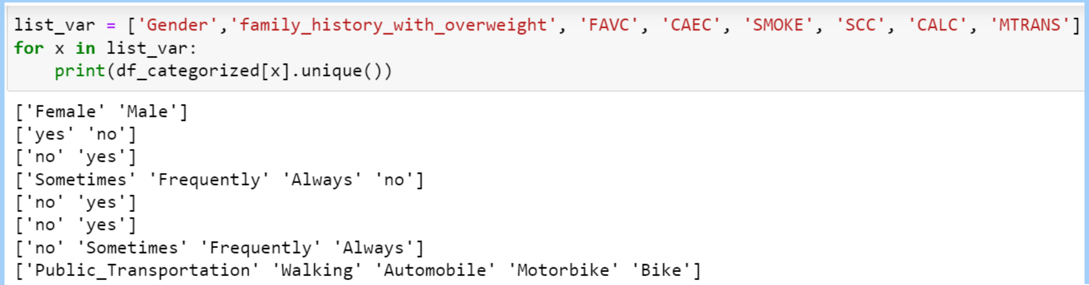
In this part we will see :

- Data cleaning : part 2
- Data processing
- Creation and selection of models.



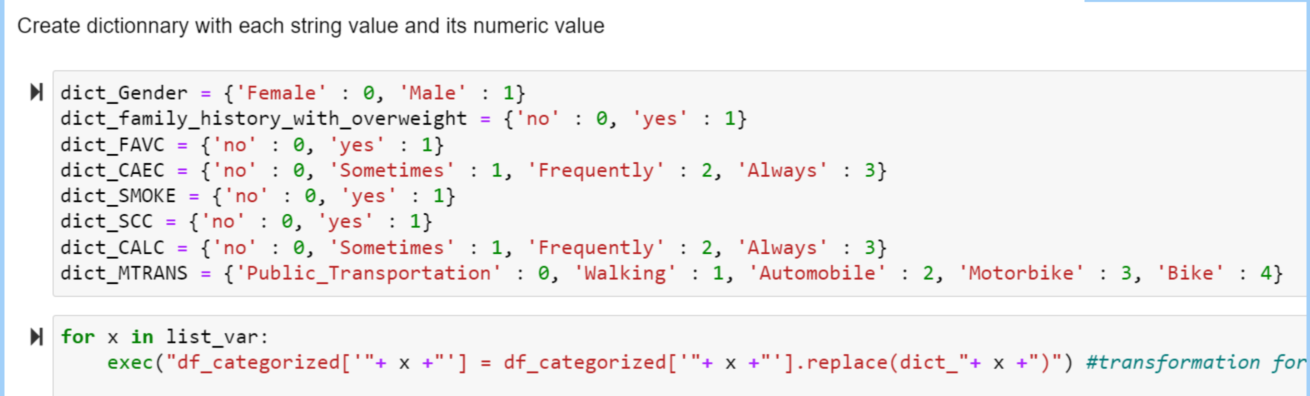
Data cleaning : part 2

Now, we can convert all our string variables into numeric ones (through categories): Let's first check what are the different possibilities for each string variables (Gender, family_history_with_overweight, 'FAVC', CAEC, SMOKE, SCC, CALC, MTRANS, NObesitydad)



```
list_var = ['Gender','family_history_with_overweight', 'FAVC', 'CAEC', 'SMOKE', 'SCC', 'CALC', 'MTRANS']
for x in list_var:
    print(df_categorized[x].unique())
```

```
['Female' 'Male']
['yes' 'no']
['no' 'yes']
['Sometimes' 'Frequently' 'Always' 'no']
['no' 'yes']
['no' 'yes']
['no' 'Sometimes' 'Frequently' 'Always']
['Public_Transportation' 'Walking' 'Automobile' 'Motorbike' 'Bike']
```



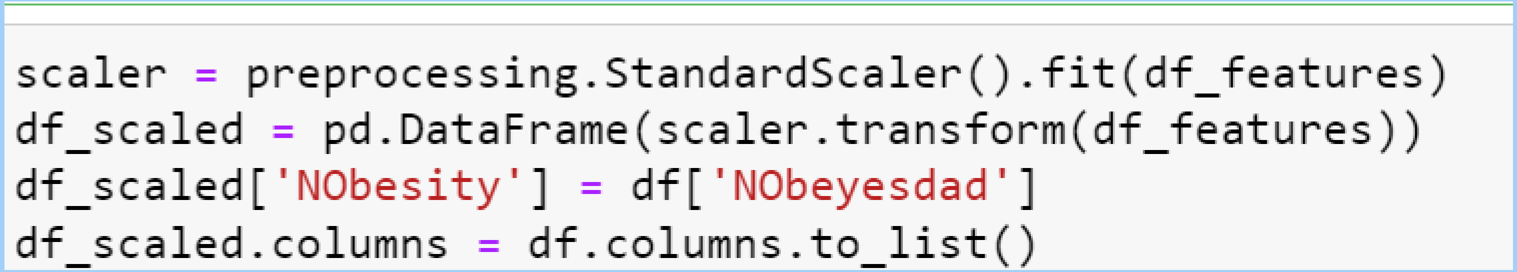
Create dictionnary with each string value and its numeric value

```
dict_Gender = {'Female' : 0, 'Male' : 1}
dict_family_history_with_overweight = {'no' : 0, 'yes' : 1}
dict_FAVC = {'no' : 0, 'yes' : 1}
dict_CAEC = {'no' : 0, 'Sometimes' : 1, 'Frequently' : 2, 'Always' : 3}
dict_SMOKE = {'no' : 0, 'yes' : 1}
dict_SCC = {'no' : 0, 'yes' : 1}
dict_CALC = {'no' : 0, 'Sometimes' : 1, 'Frequently' : 2, 'Always' : 3}
dict_MTRANS = {'Public_Transportation' : 0, 'Walking' : 1, 'Automobile' : 2, 'Motorbike' : 3, 'Bike' : 4}
```

```
for x in list_var:
    exec("df_categorized['"+ x +"'] = df_categorized['"+ x +"'].replace(dict_"+ x +")") #transformation for
```

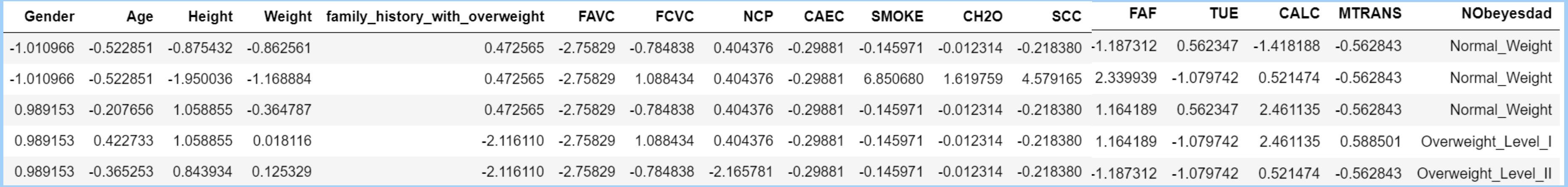
Data processing

We perform a standardization of the dataset in order to improve the performance of our models



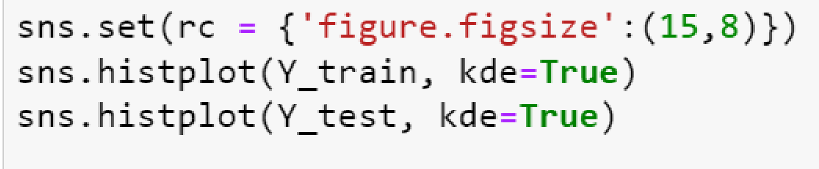
```
scaler = preprocessing.StandardScaler().fit(df_features)
df_scaled = pd.DataFrame(scaler.transform(df_features))
df_scaled['NObesity'] = df['NObeyesdad']
df_scaled.columns = df.columns.to_list()
```

Final data set :

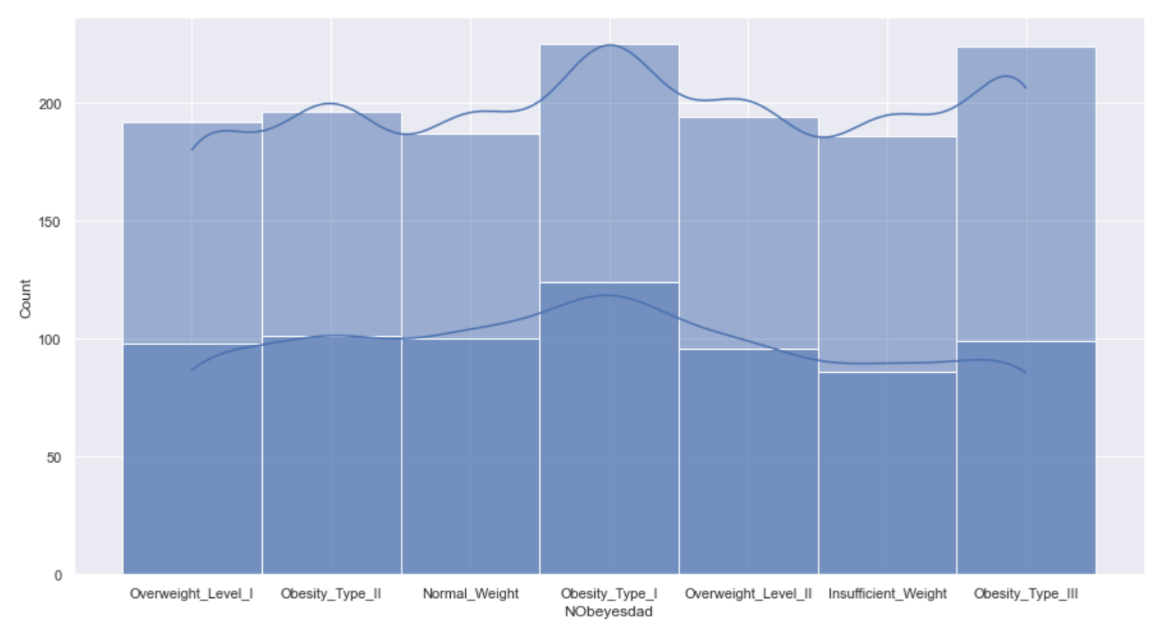


Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
-1.010966	-0.522851	-0.875432	-0.862561	0.472565	-2.75829	-0.784838	0.404376	-0.29881	-0.145971	-0.012314	-0.218380	-1.187312	0.562347	-1.418188	-0.562843	Normal_Weight
-1.010966	-0.522851	-1.950036	-1.168884	0.472565	-2.75829	1.088434	0.404376	-0.29881	6.850680	1.619759	4.579165	2.339939	-1.079742	0.521474	-0.562843	Normal_Weight
0.989153	-0.207656	1.058855	-0.364787	0.472565	-2.75829	-0.784838	0.404376	-0.29881	-0.145971	-0.012314	-0.218380	1.164189	0.562347	2.461135	-0.562843	Normal_Weight
0.989153	0.422733	1.058855	0.018116	-2.116110	-2.75829	1.088434	0.404376	-0.29881	-0.145971	-0.012314	-0.218380	1.164189	-1.079742	2.461135	0.588501	Overweight_Level_I
0.989153	-0.365253	0.843934	0.125329	-2.116110	-2.75829	-0.784838	-2.165781	-0.29881	-0.145971	-0.012314	-0.218380	-1.187312	-1.079742	0.521474	-0.562843	Overweight_Level_II

Train and test set



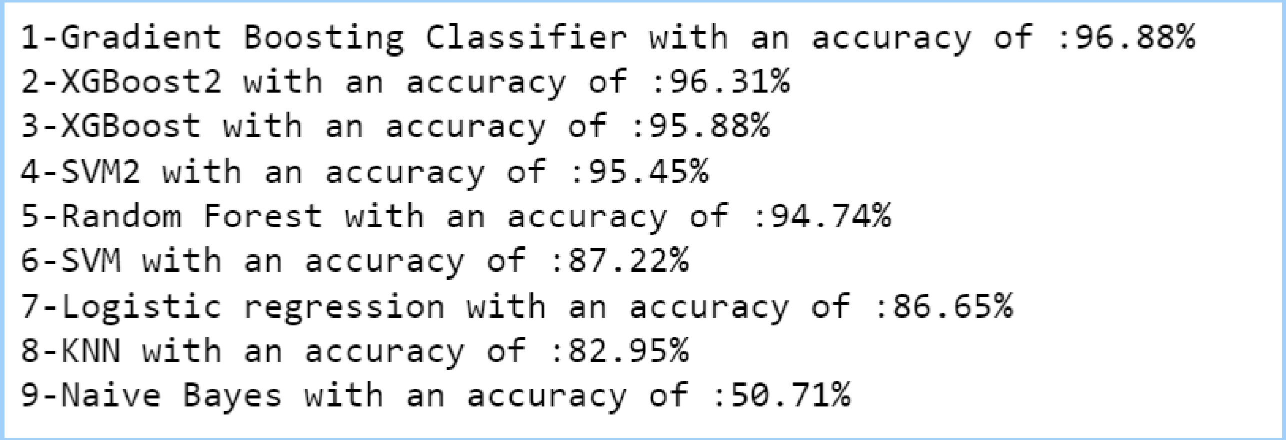
```
sns.set(rc = {'figure.figsize':(15,8)})
sns.histplot(Y_train, kde=True)
sns.histplot(Y_test, kde=True)
```



The split between train and test is correct, because it is still well balanced.
We choose to do 66% of the data for the train and 33% for the test.

Models

Here is the ranking of the best prediction models we use on the data set :

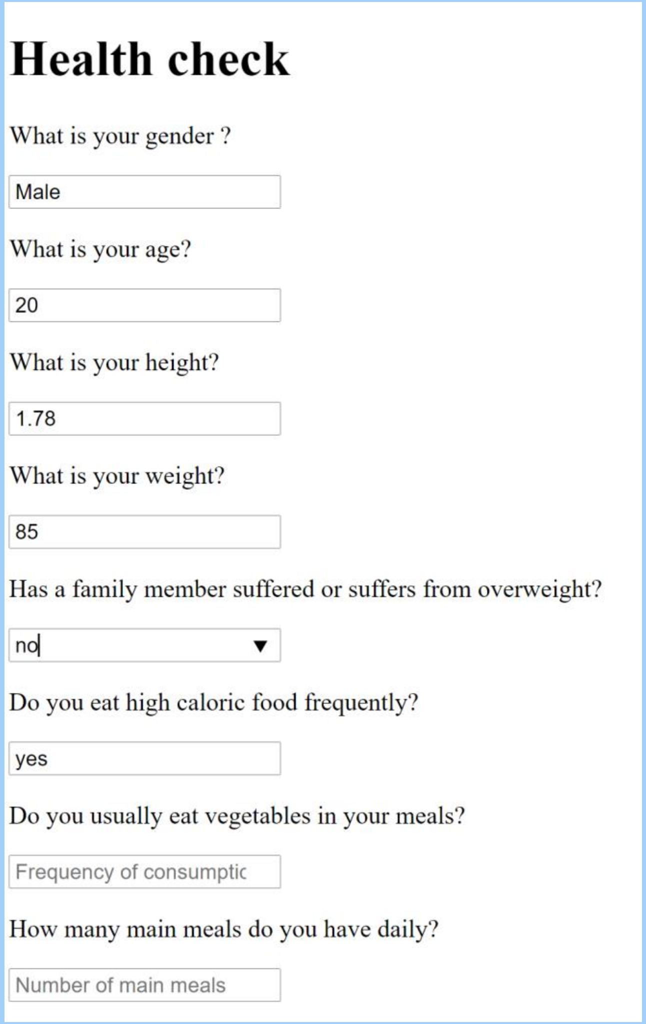


```
1-Gradient Boosting Classifier with an accuracy of :96.88%
2-XGBoost2 with an accuracy of :96.31%
3-XGBoost with an accuracy of :95.88%
4-SVM2 with an accuracy of :95.45%
5-Random Forest with an accuracy of :94.74%
6-SVM with an accuracy of :87.22%
7-Logistic regression with an accuracy of :86.65%
8-KNN with an accuracy of :82.95%
9-Naive Bayes with an accuracy of :50.71%
```

So for our model and API we will use the gradient boosting classifier model because it is the one that has the best performance.

The API

The API is interactive you can select the correct answer on a list for each questions then push the button « Calculate » to determine your level of obesity according to science.
You can refresh the page to do another try.



Health check

What is your gender ?

Male

What is your age?

20

What is your height?

1.78

What is your weight?

85

Has a family member suffered or suffers from overweight?

no ▼

Do you eat high caloric food frequently?

yes

Do you usually eat vegetables in your meals?

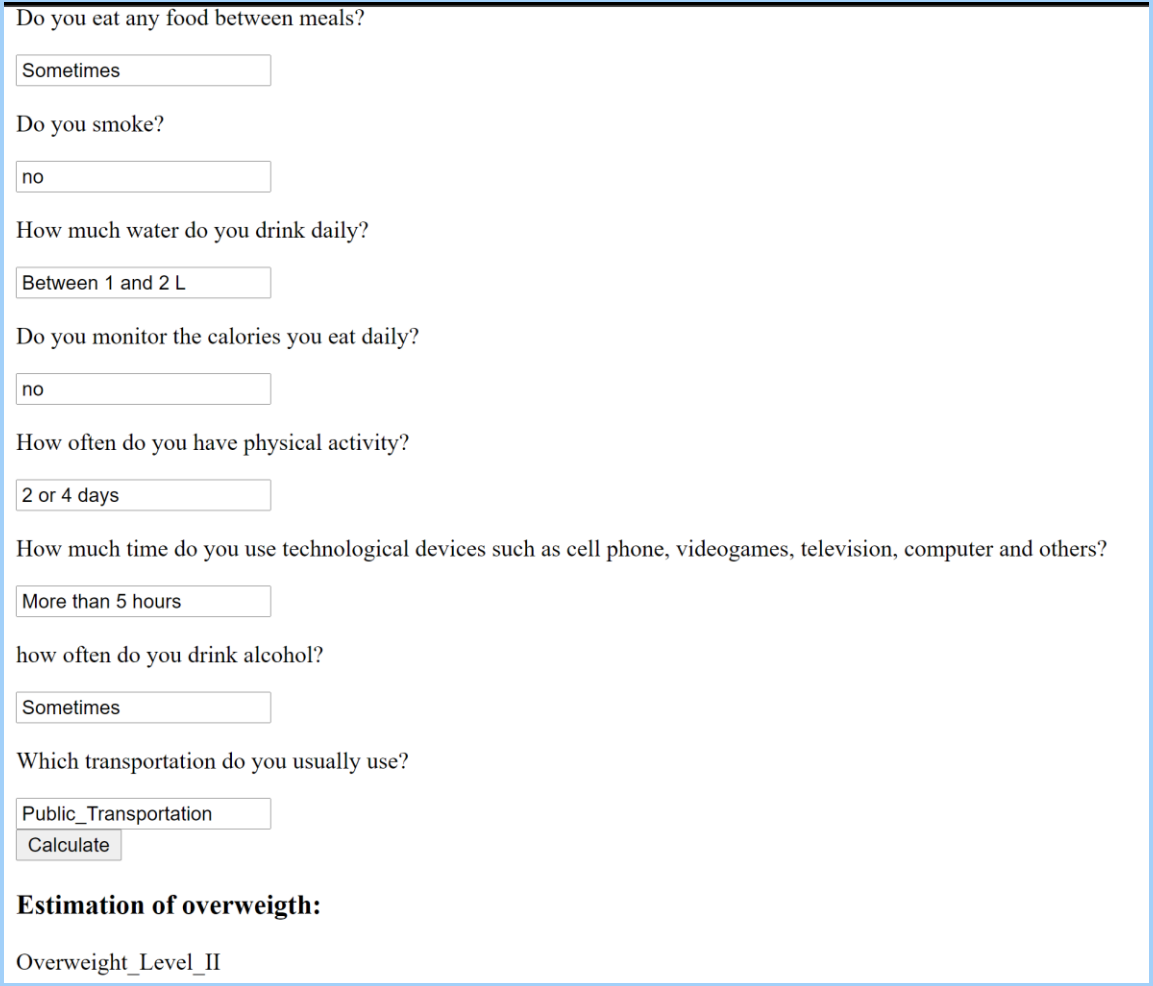
Frequency of consumptic

How many main meals do you have daily?

Number of main meals

The API

The API is interactive you can select the correct answer on a list for each questions then push the button « Calculate » to determine your level of obesity according to science.
You can refresh the page to do another try.



Do you eat any food between meals?

Sometimes

Do you smoke?

no

How much water do you drink daily?

Between 1 and 2 L

Do you monitor the calories you eat daily?

no

How often do you have physical activity?

2 or 4 days

How much time do you use technological devices such as cell phone, videogames, television, computer and others?

More than 5 hours

how often do you drink alcohol?

Sometimes

Which transportation do you usually use?

Public_Transportation

Estimation of overweigth:

Overweight_Level_II

Thank you !