Estimation of obesity based on eating habits and physical condition

Mathilda Charoy and Lila Allanic
DIA1

Introduction

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

There are 6 types of obesity : insufficient weight, normal weight, overweight level I, overweight level II, obesity type I and obesity type II.

The dataset

The dataset we used is based on informations about eating habits and physical condition of individuals from Colombia, Peru and Mexico.

The dataset contains 16 variables :
Gender, age, height, weight, family history with overweight, attributes related t eating habits (6) and attributes related to physical condition (4).

With one target:
NObeyesdad, equivalent to NObesity

Dataset description

Description available at : sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub

Columns	Related question	Possible values
Gender	Gender	'Male' 'Female'
Age	Age	Int
Height	Height	Float in meters
Weight	Weight	Float in kg
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'

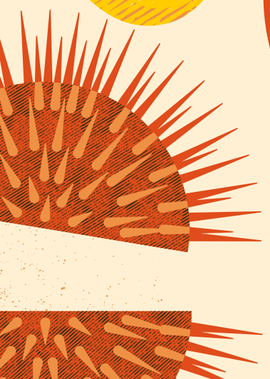
Dataset description

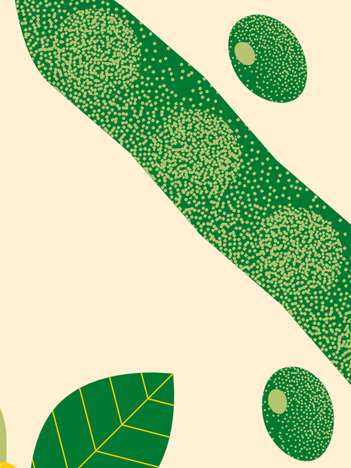
Columns	Related question	Possible values
NCP	How many main meals do you have daily ?	Int
CAEC	Do you eat any food between meals ?	'No' 'Sometimes' 'Frequently' 'Always'
SMOKE	Do you smoke ?	'Yes' 'No'
CH2O	How much water do you drink daily ?	Float in L
SCC	Do you monitor the calories you eat daily ?	'Yes' 'No'
FAF	How often do you have physical activity weekly ?	Int
TUE	How much time do you use technological devices daily ?	Float in hours

Dataset description

Columns	Related question	Possible values
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



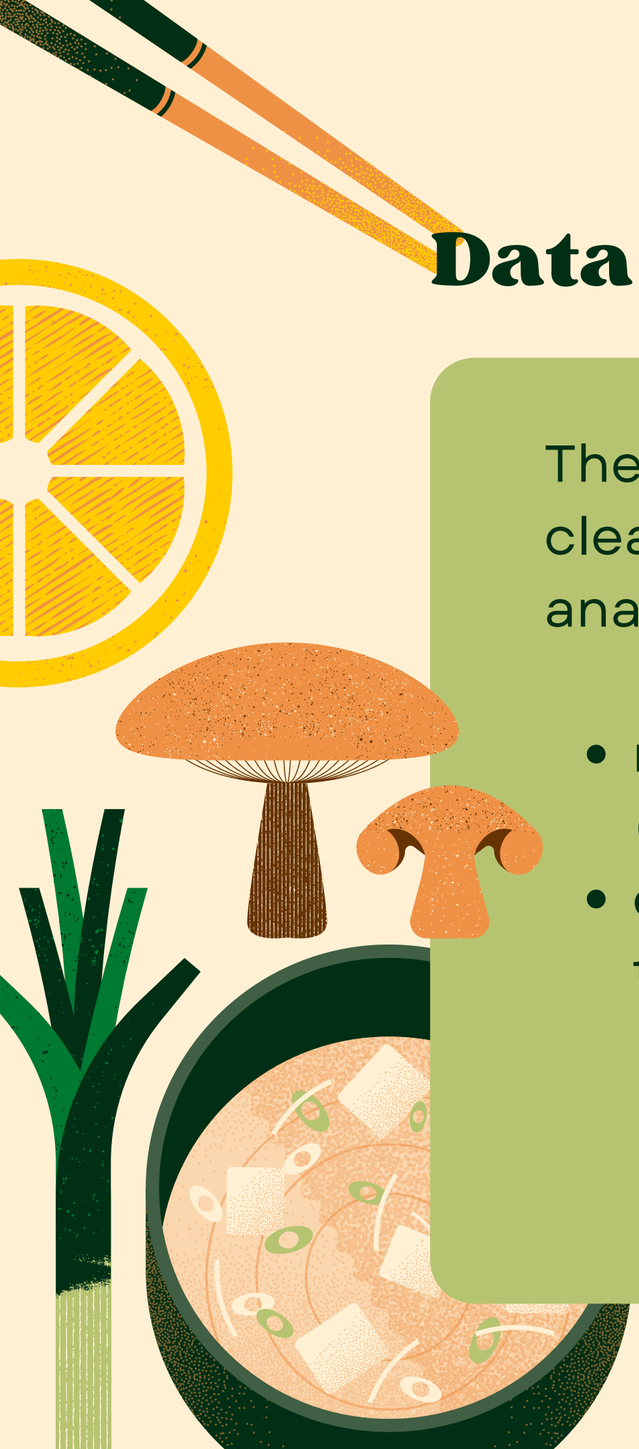


This part will be divided in two :

Data cleaning Part I

Data Analysis

Data cleaning Part I



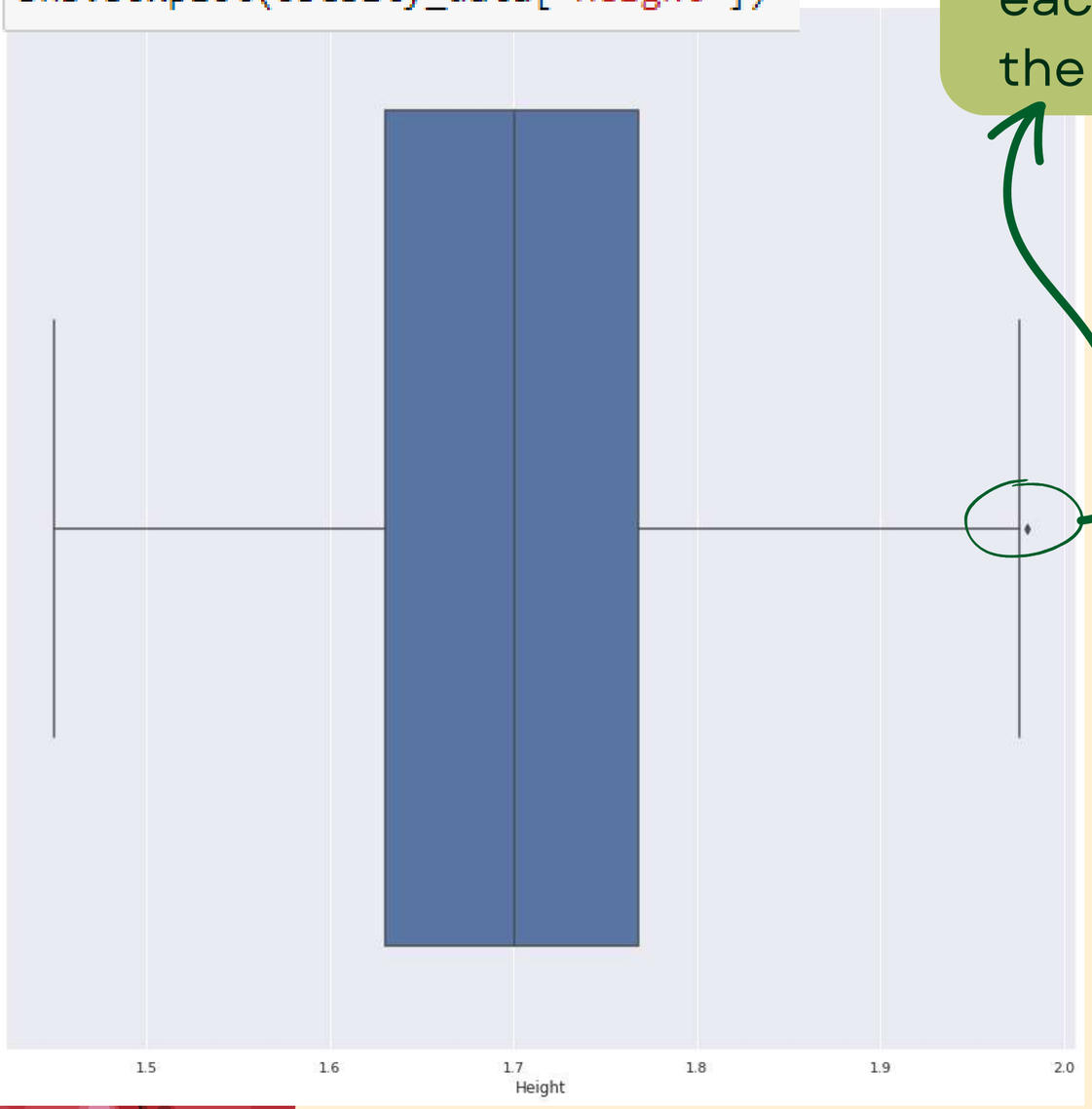
The dataset didn't need much cleaning to be ready for data analysis.

- no Null variables (df.dropna())
- conversion of variable types

Checking for outliers :

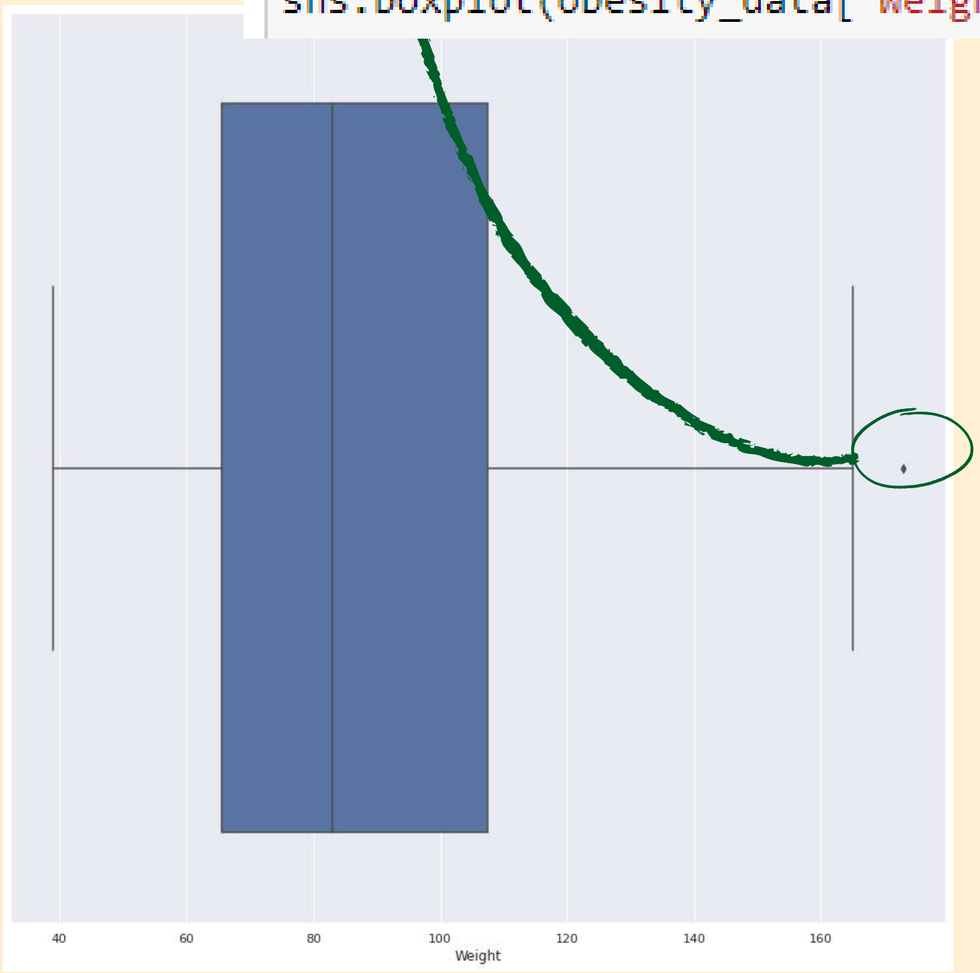
For non-categorical variables, we need to check if there is any outliers (variables too low or to high in value). There are 3 concerned variables : **Age, Height and Weigh**t. Age is not a problem, because the study has been done with subjects between 14 and 61 years.

```
sns.boxplot(obesity_data['Height'])
```



As you can see on the right side of each plot, there is a dot attesting the presence of an outlier.

```
sns.boxplot(obesity_data['Weight'])
```



General view of quantitative data :

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000
mean	24.312600	1.701677	86.586058	2.419043	2.685628	2.008011	1.010298	0.657866
std	6.345968	0.093305	26.191172	0.533927	0.778039	0.612953	0.850592	0.608927
min	14.000000	1.450000	39.000000	1.000000	1.000000	1.000000	0.000000	0.000000
25%	19.947192	1.630000	65.473343	2.000000	2.658738	1.584812	0.124505	0.000000
50%	22.777890	1.700499	83.000000	2.385502	3.000000	2.000000	1.000000	0.625350
75%	26.000000	1.768464	107.430682	3.000000	3.000000	2.477420	1.666678	1.000000
max	61.000000	1.980000	173.000000	3.000000	4.000000	3.000000	3.000000	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 median age is 27 years old**.

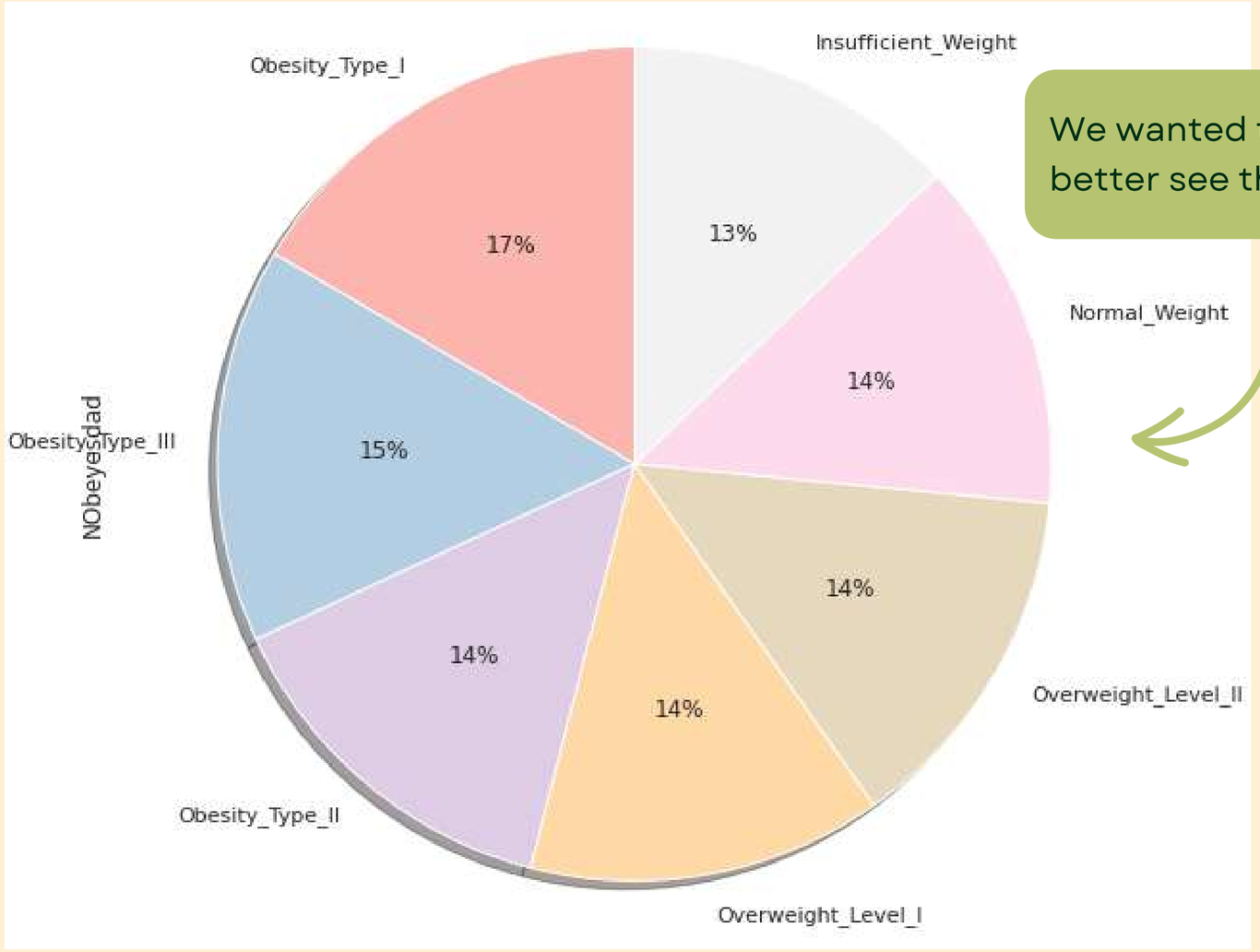
Data Analysis

This study will be divided in two :

The features of the population

The lifestyle of the population (consumption)

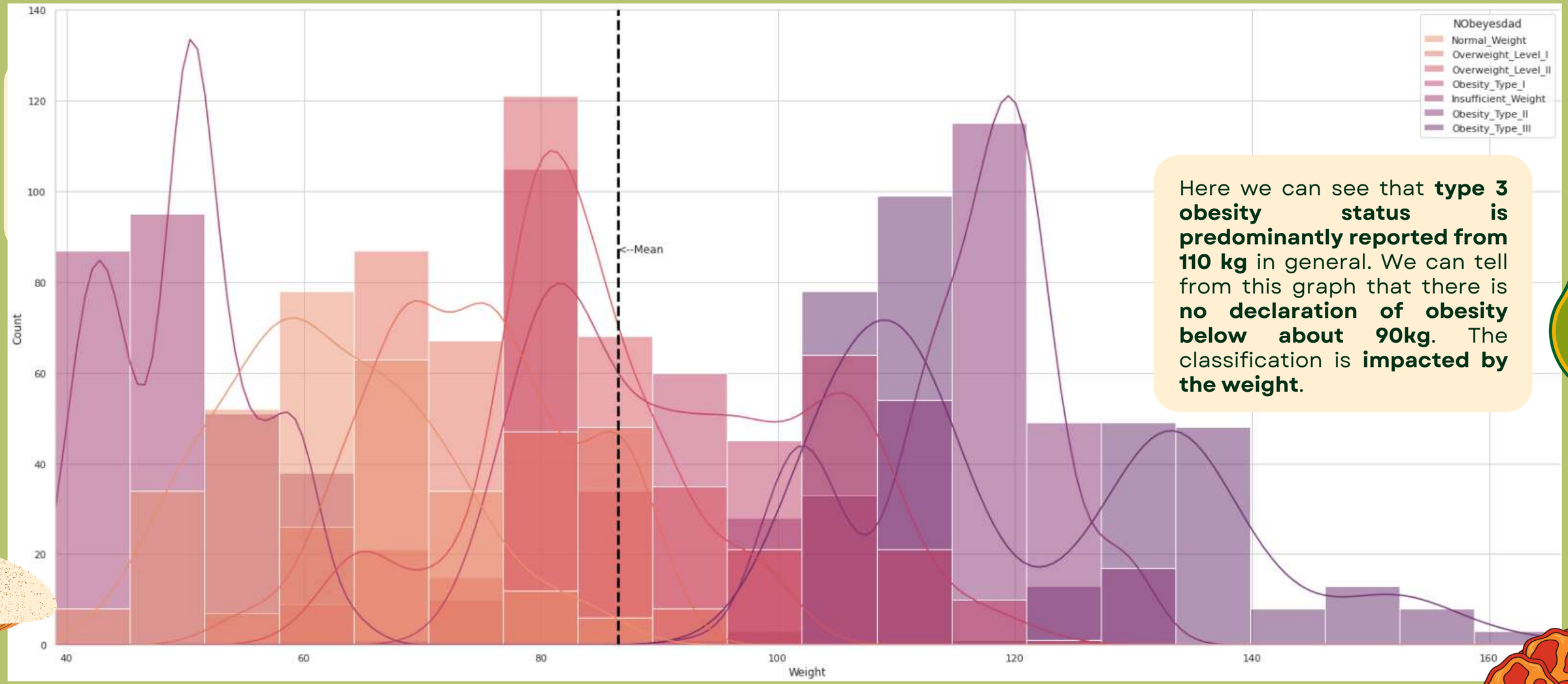
Features of the population



We wanted to have a global view to better see the population's repartition.

Weight

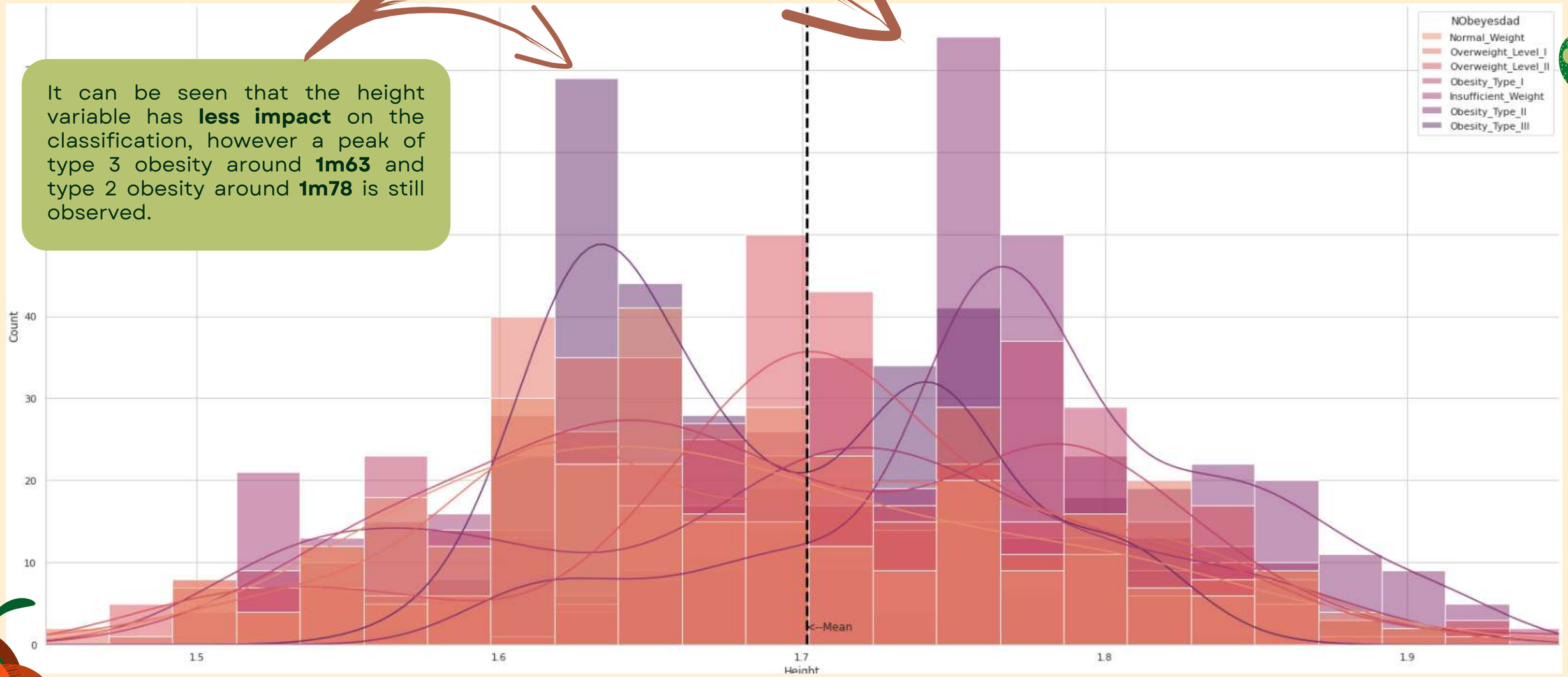
Then we looked at the variables h**eight, weight, age, gender** and **family history** that we thought were the most important (by plotting the histograms that represent the status according to the variable)



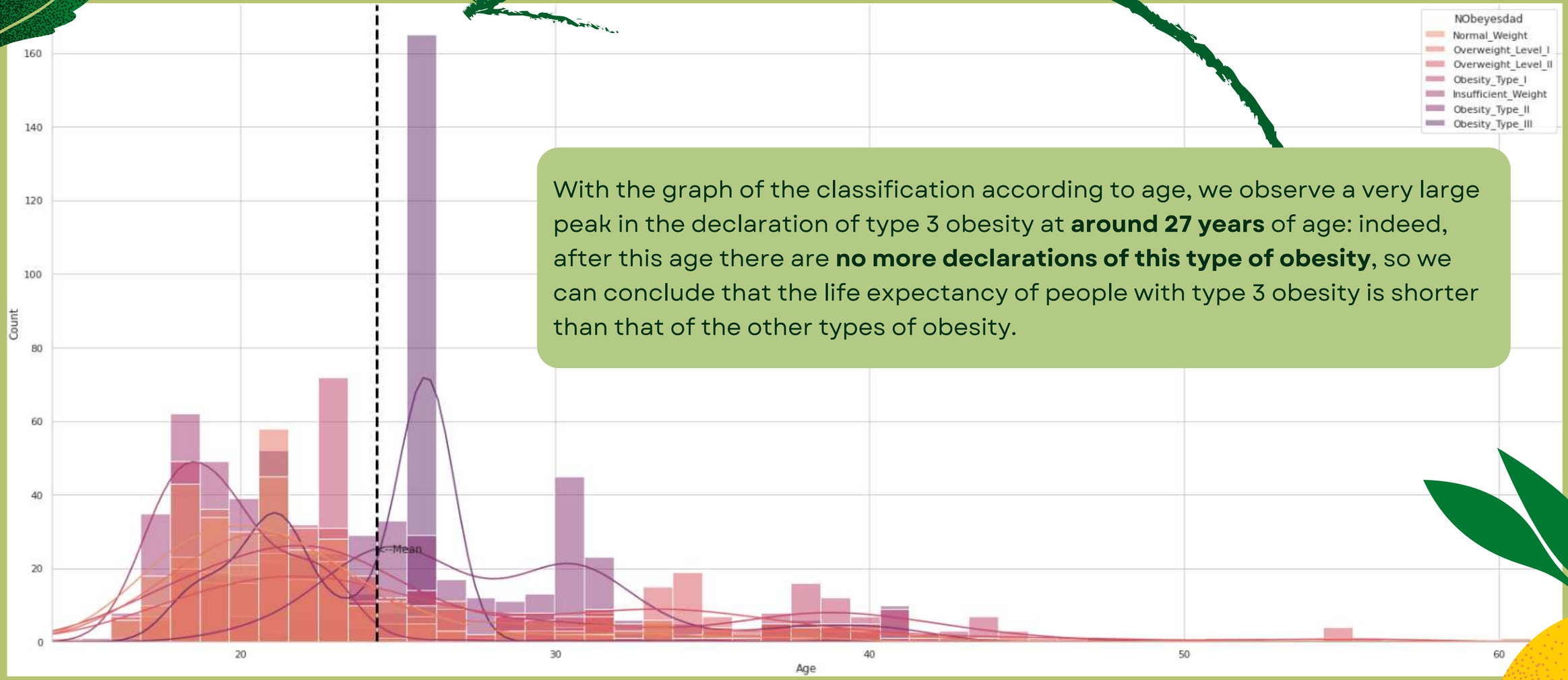
Here we can see that **type 3 obesity status is predominantly reported from 110 kg** in general. We can tell from this graph that there is **no declaration of obesity below about 90kg**. The classification is **impacted by the weight**.

Height

It can be seen that the height variable has **less impact** on the classification, however a peak of type 3 obesity around **1m63** and type 2 obesity around **1m78** is still observed.

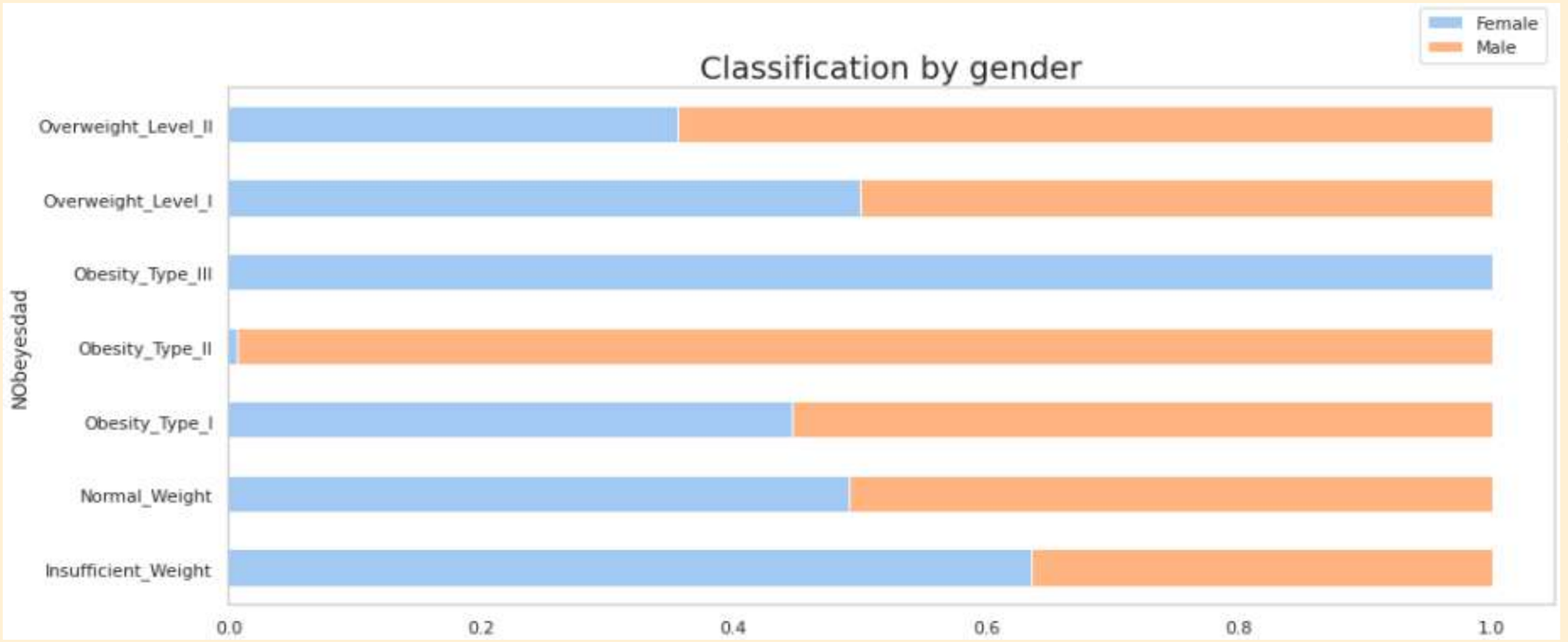


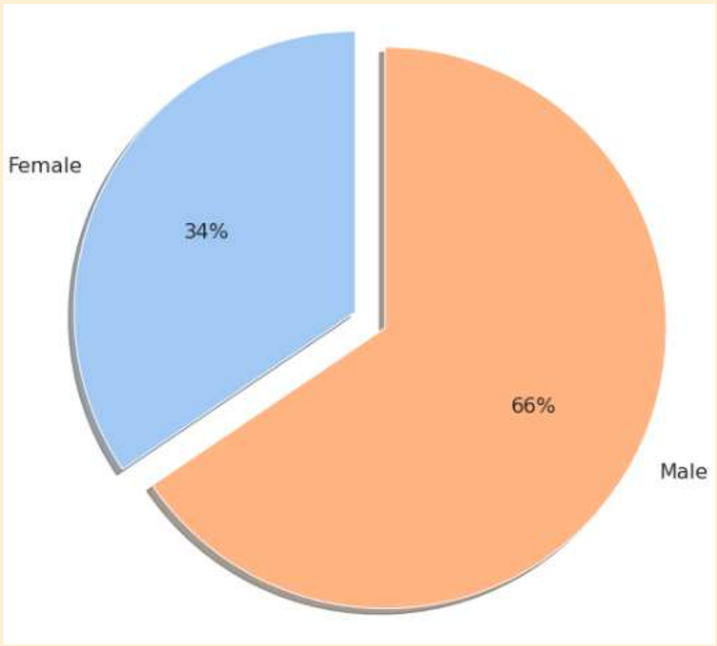
Age



With the graph of the classification according to age, we observe a very large peak in the declaration of type 3 obesity at **around 27 years** of age: indeed, after this age there are **no more declarations of this type of obesity**, so we can conclude that the life expectancy of people with type 3 obesity is shorter than that of the other types of obesity.

Gender

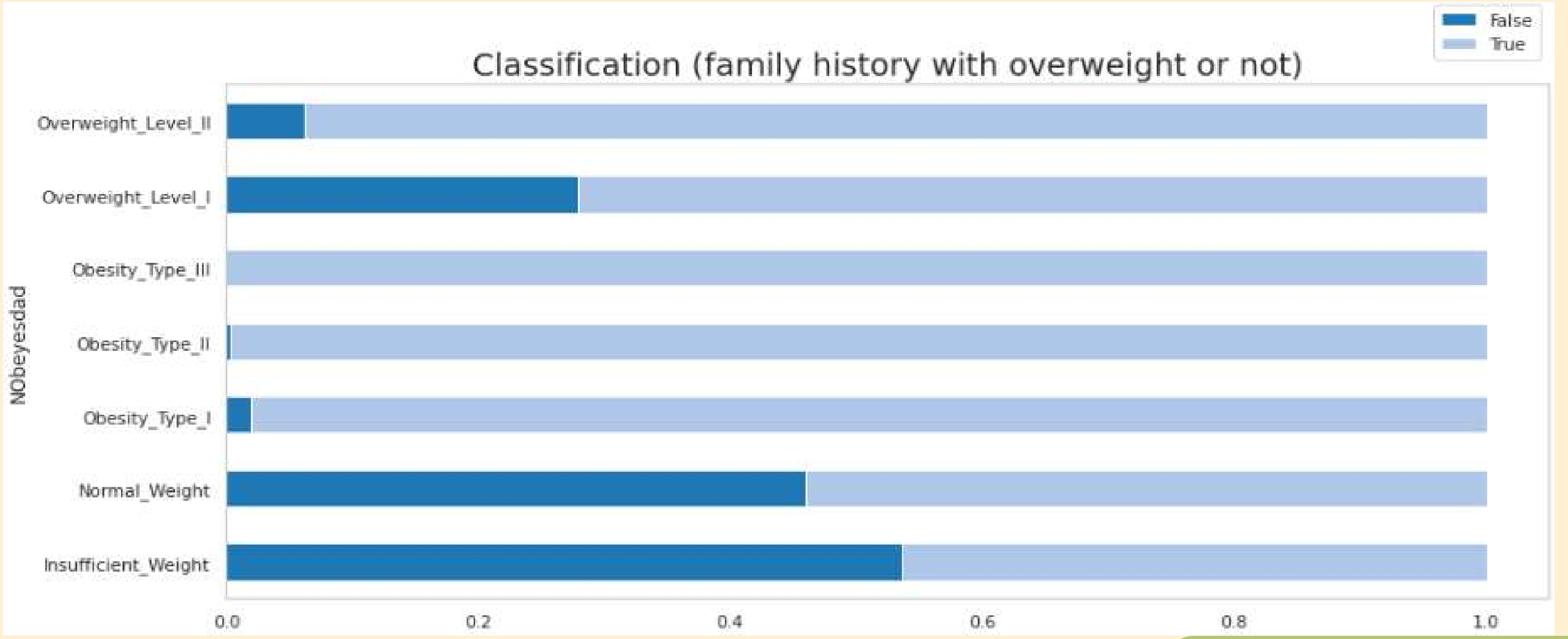




Here, we can see that there is more cases of obesity and overweight in men, excepted for the type III. And we can tell that the women are more affected by the insufficient weight.

66% of the population affected by obesity, all types combined, are men, which is twice as many as the female (**34%**).

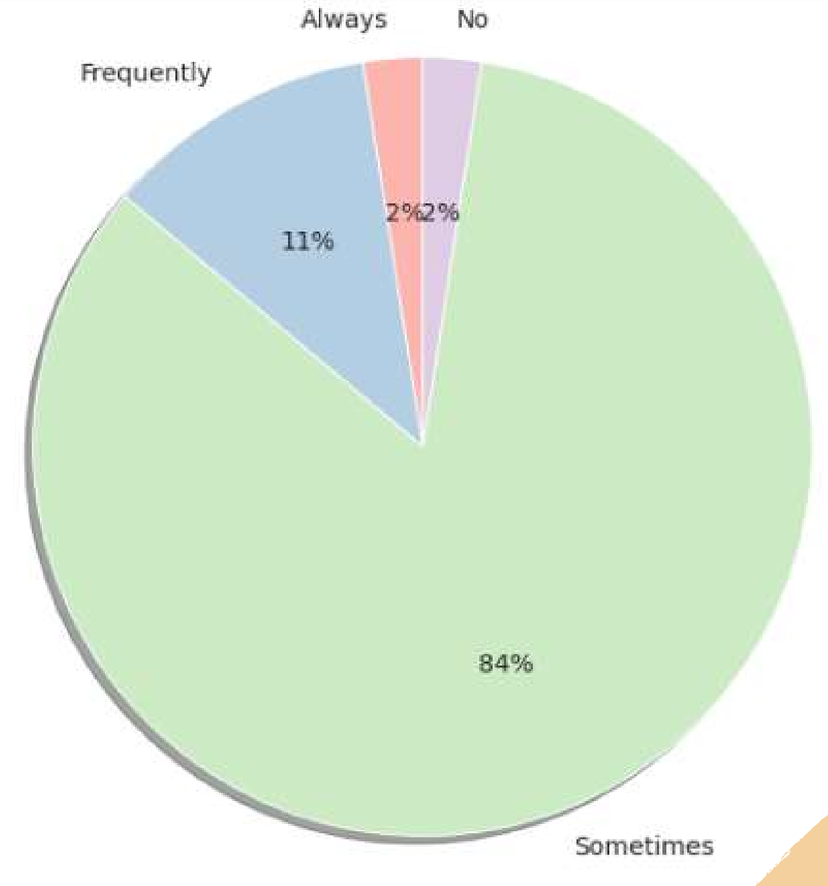
Family history

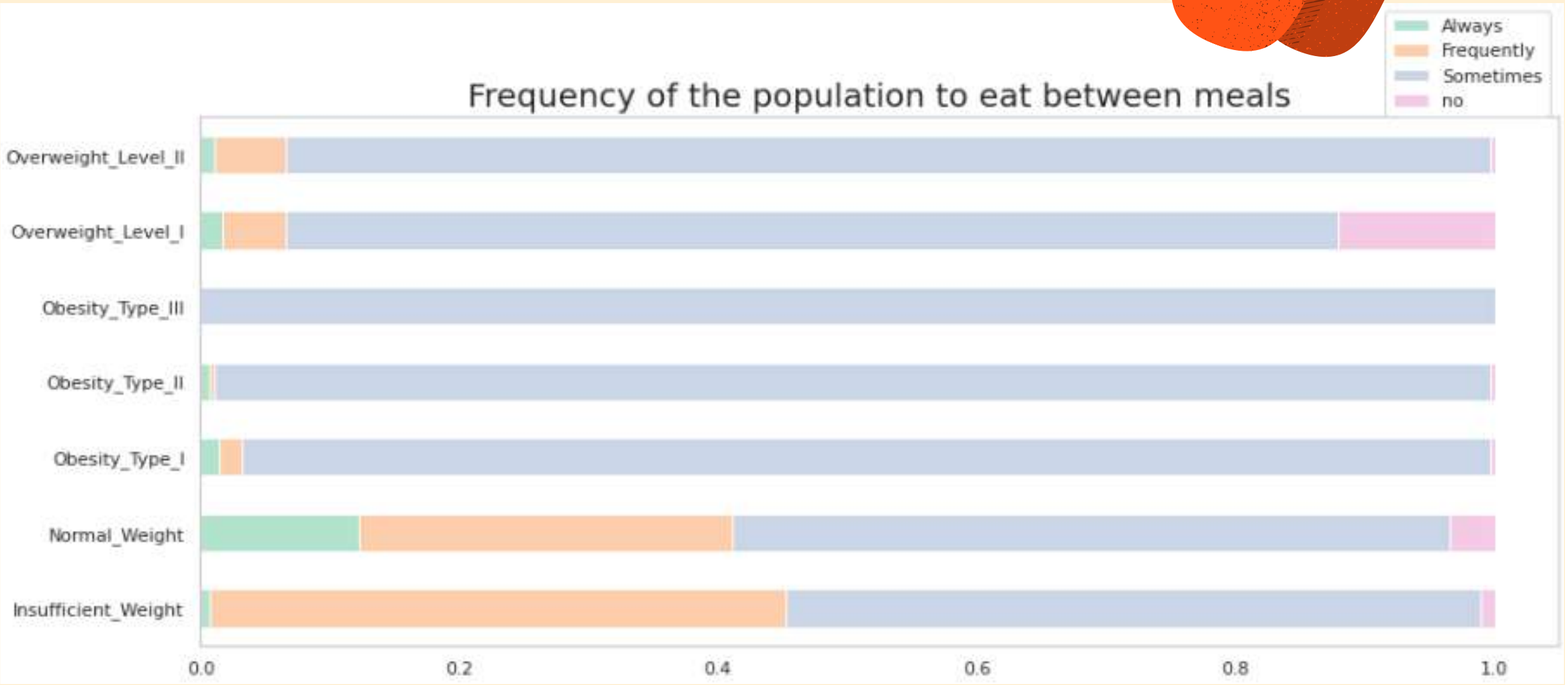


This graph shows that **the higher the level of obesity, the more likely it is that the subject has a family history of overweight.** It is therefore a determining factor.

Lifestyle and consumption

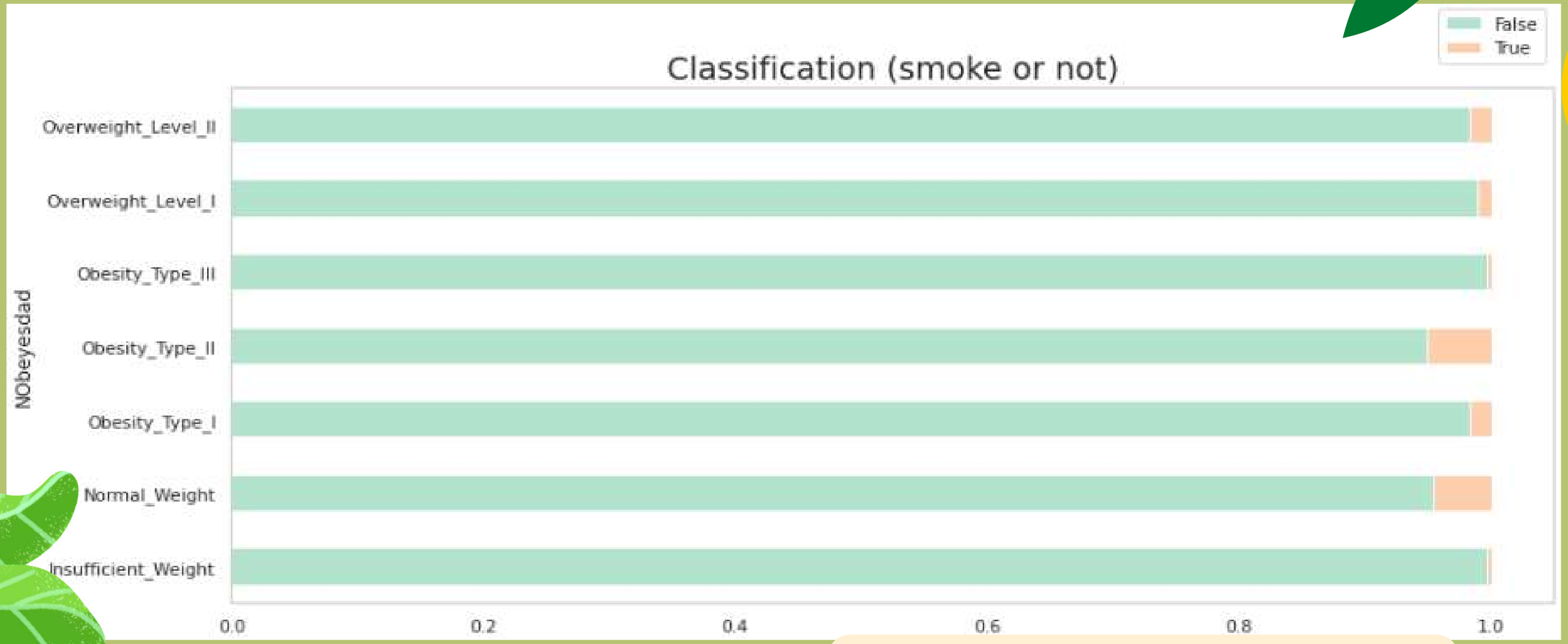
How ofter do you eat between meals ?





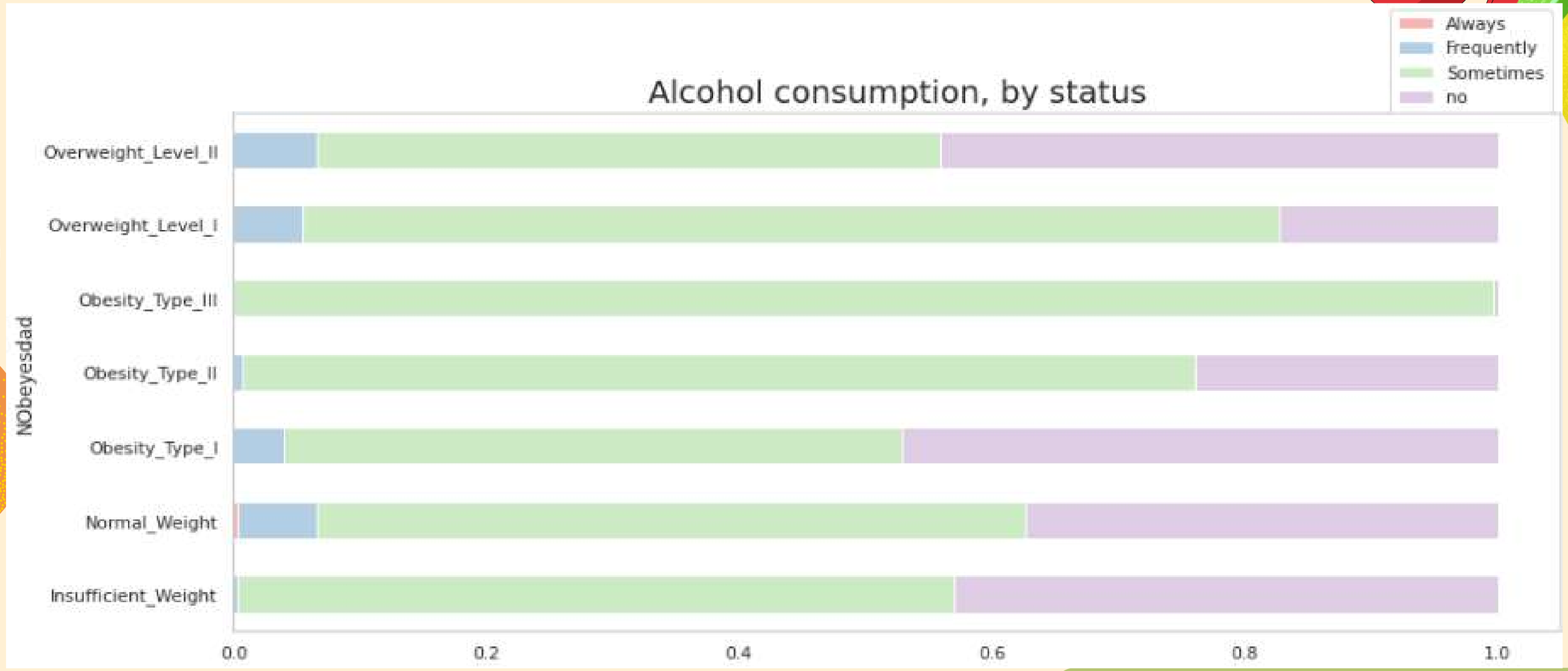
We notice that, contrary to what we might think, **the people with the highest obesity status are the ones who eat the least between meals**, and that it is the ones in normal or insufficient weight who eat the most.

Do you smoke ?



We notice in this graph that there are **very few individuals who smoke, regardless of status**. It is therefore not a factor to be taken into consideration.

How often do you drink alcohol ?

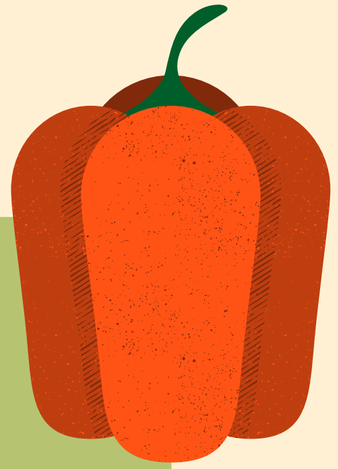


It can be deduced from this graph that **the more important the type of overweight or obesity, the less frequently people consume**. The proof is in the type III.

What type of transport do you use ?

We can observe that the **most taken transport is the public transport** and that the car comes after. The people that are in the obesity type III case only take the public transport and don't use the physical transports.

Modelization

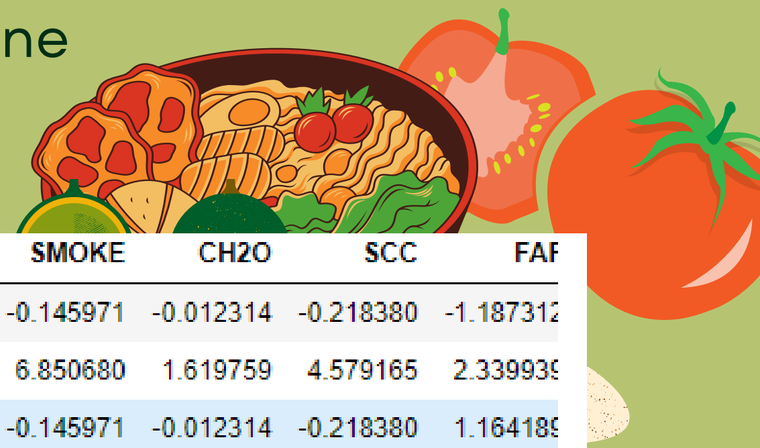


In this part we will see :

- **Data cleaning Part II and data processing**
- **Splitting the data into train and test datasets**
- **Creation and selection of models.**

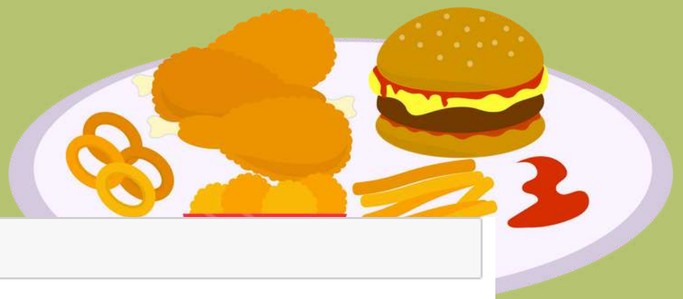
Data cleaning and processing

Now, we can convert all our string variables into numeric ones (through categories). We also perform a standardization of the dataset in order to improve the performance of our models.
We end up with this dataset, ready for machine learning !



	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF
0	-1.010966	-0.522851	-0.875432	-0.862561	0.472565	-2.75829	-0.784838	0.404376	-0.356879	-0.145971	-0.012314	-0.218380	-1.187312
1	-1.010966	-0.522851	-1.950036	-1.168884	0.472565	-2.75829	1.088434	0.404376	-0.356879	6.850680	1.619759	4.579165	2.339939
2	0.989153	-0.207656	1.058855	-0.364787	0.472565	-2.75829	-0.784838	0.404376	-0.356879	-0.145971	-0.012314	-0.218380	1.164189
3	0.989153	0.422733	1.058855	0.018116	-2.116110	-2.75829	1.088434	0.404376	-0.356879	-0.145971	-0.012314	-0.218380	1.164189
4	0.989153	-0.365253	0.843934	0.125329	-2.116110	-2.75829	-0.784838	-2.165781	-0.356879	-0.145971	-0.012314	-0.218380	-1.187312

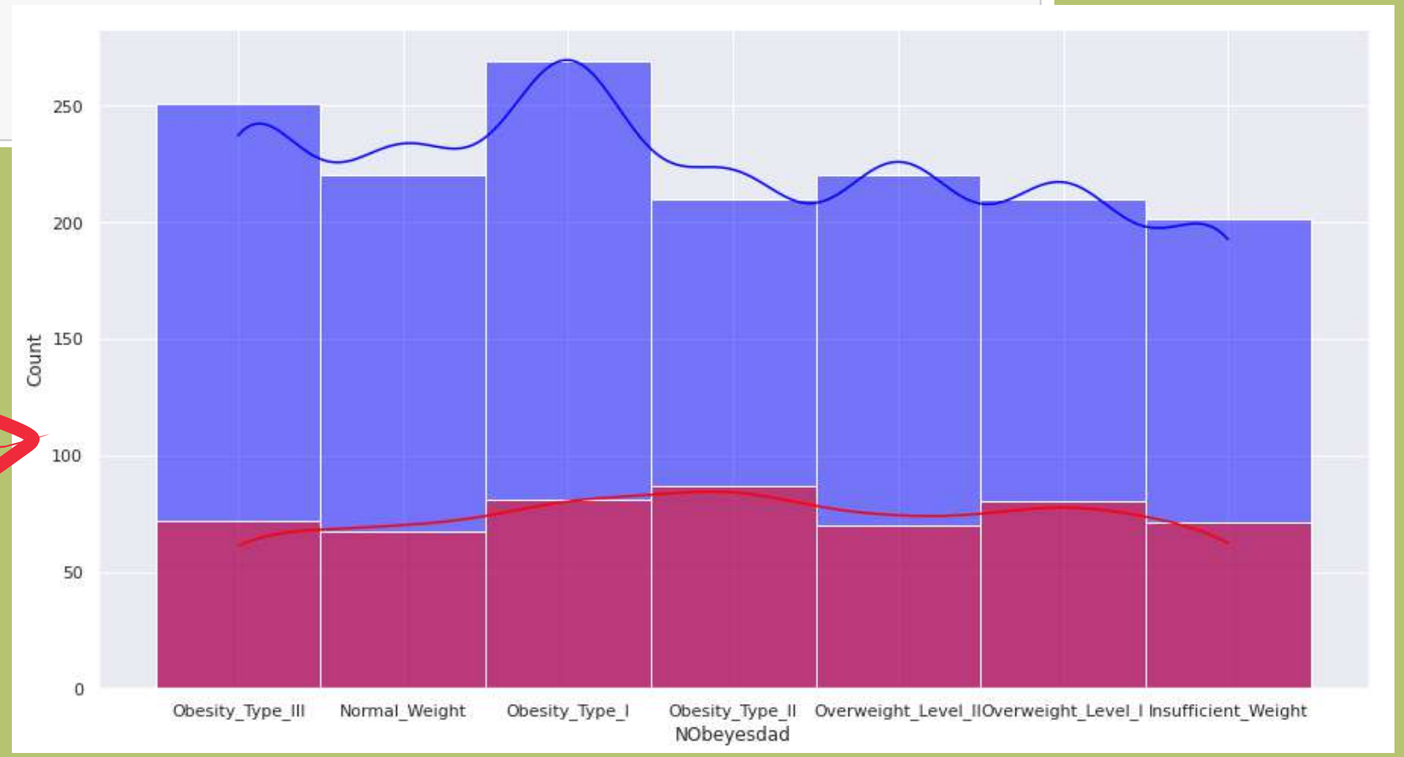
Splitting the data into train and test datasets



```
train_test_dataset = obesity_data_scaled.drop(columns="NObeyesdad")
```

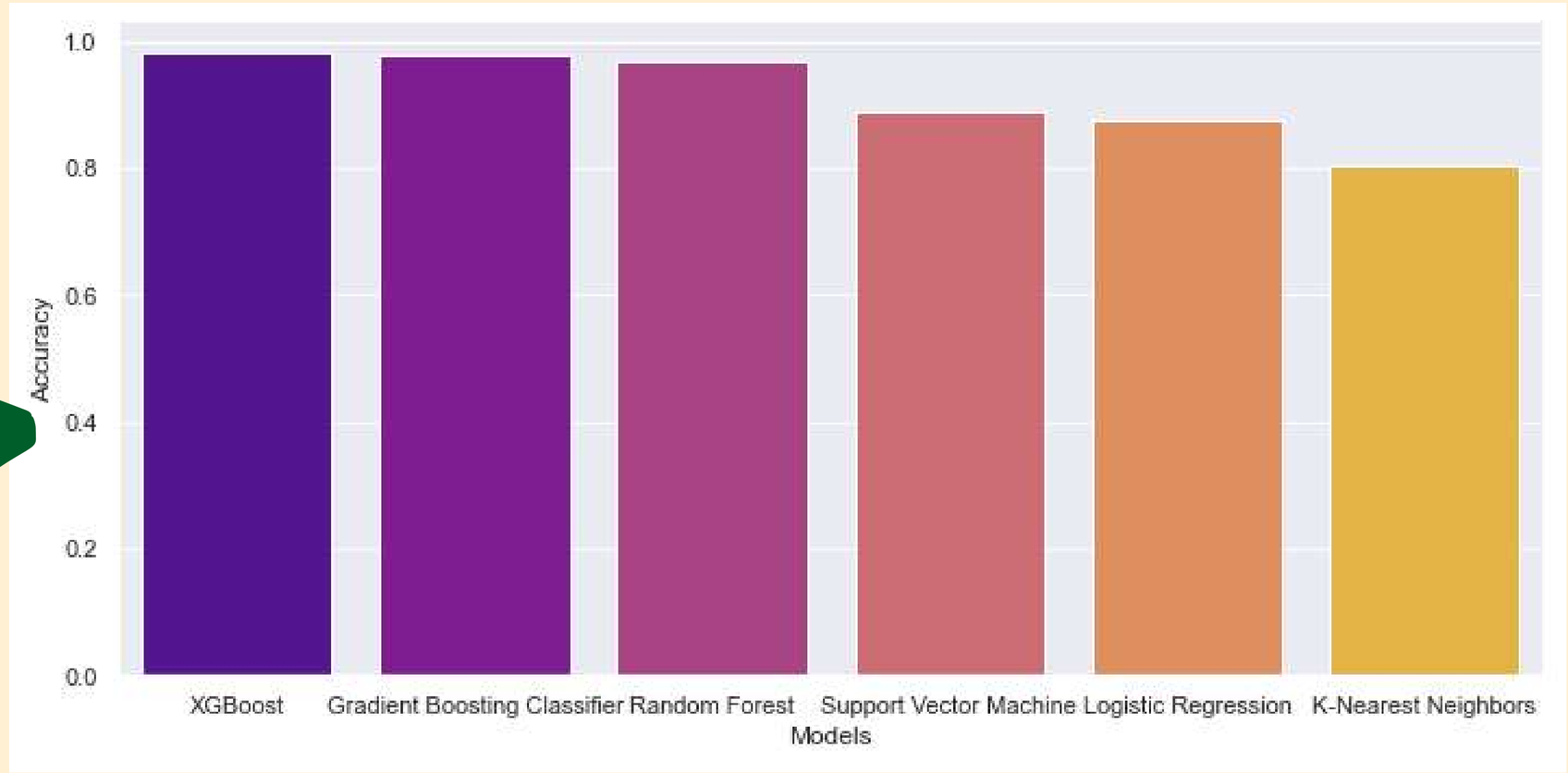
```
X_train, X_test, Y_train, Y_test = train_test_split(train_test_dataset, obesity_data['NObeyesdad'])
```

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



Ranking of the models

	Models	Accuracy
0	XGBoost	0.981061
1	Gradient Boosting Classifier	0.979167
2	Random Forest	0.965909
3	Support Vector Machine	0.888258
4	Logistic Regression	0.875000
5	K-Nearest Neighbors	0.803030



The API

using Dash

The API is interactive, a bunch of questions is asked to the user and he can select or type the correct answer.
When pushing the butter "Compute", our IA trained on the Obesity dataset will determine their level of obesity and display it on the screen.

A quick look at our API

ESTIMATION OF OBESITY LEVELS

We are going to ask you some questions to try and define the odds of you being obese and the level of obesity that touches or will touch you. Please fill in the following informations :

Gender

Select...

Age

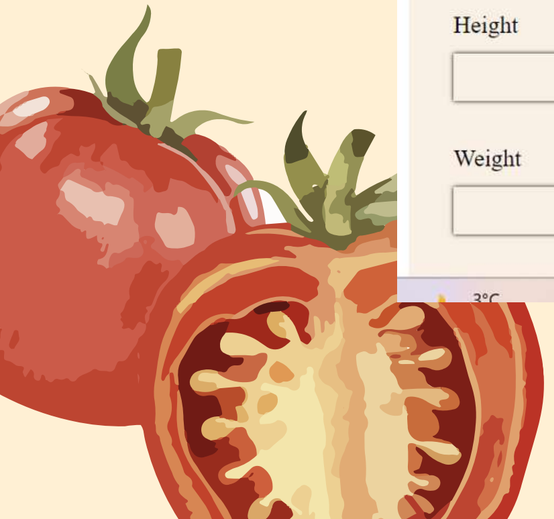
Number corresponding

Height

Ex : 1.72

Weight

In kg



How the answer is shown on the API

Thank you for listening !

Mathilda Charoy and Lila Allanic
DIA1