ESiLV
LÉONARD
DE VINCI

GRADUATE SCHOOL OF
ENGINEERING
PARIS-LA DÉFENSE

ESILV- Python for Data Analysis - Project

Written by Jihane Oul Ali and Nada Matrouf

Content

- Data Introduction
- Data Visualization
- Data Preparation
- Modelisation and Validation
- Conclusion
- References

1. Data Introduction

Abstract

The data represents the estimation of obesity levels in individuals from Mexico, Peru and Columbia. The estimation is based on their eating habits and physical condition. The data contains 17 attributed with 2111 instances.

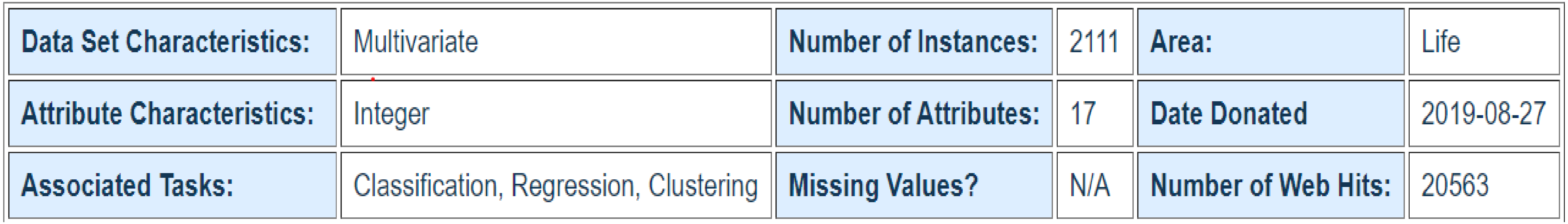
The target is labeled with the following values :

- *Insufficient Weight*
- *Normal Weight*
- *Overweight Level I*
- *Overweight Level II*
- *Obesity Type I*
- *Obesity Type II*
- *Obesity Type III*

1. Data Introduction

Link:

https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+



Data Set Characteristics:	Multivariate	**Number of Instances:**	2111	**Area:**	Life
Attribute Characteristics:	Integer	**Number of Attributes:**	17	**Date Donated**	2019-08-27
Associated Tasks:	Classification, Regression, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	20563

1. Data Introduction

The attributes of the dataset and their input are based on the following:

<table>
<tr><td><b>Gender</b></td><td><i>Male |Female</i></td><td><b>Food consumption between meals</b></td><td><i>No | Sometimes | Frequently | Always</i></td></tr>
<tr><td><b>Age</b></td><td><i>Numeric Value</i></td><td><b>Smoking</b></td><td><i>Boolean</i></td></tr>
<tr><td><b>Height</b></td><td><i>Num. Val. In meters</i></td><td><b>Water consumption</b></td><td><i>Less than 1L| Between 1 and 2L| More than 2L</i></td></tr>
<tr><td><b>Weight</b></td><td><i>Num. Val. In Kg</i></td><td><b>Are calories monitored daily</b></td><td><i>Yes | No</i></td></tr>
<tr><td><b>Family member was or is overweight</b></td><td><i>Boolean</i></td><td><b>Frequency of physical activity</b></td><td><i>I do not have | 1 or 2 days| 2 or 4 days | 4 or 5 days</i></td></tr>
<tr><td><b>Frequent consumption of high caloric food</b></td><td><i>Boolean</i></td><td><b>Time spent on technological devices</b></td><td><i>0-2 hours | 3-5 hours| more than 5 hours</i></td></tr>
<tr><td><b>Frequency of vegetable consumption</b></td><td><i>Never | Sometimes | Always</i></td><td><b>Frequency of alcohol consumption</b></td><td><i>I do not drink| sometimes| frequently |always</i></td></tr>
<tr><td><b>Number of daily meals</b></td><td><i>Between 1 and 2 | Three| More than three</i></td><td><b>Type of transportation used</b></td><td><i>Auto| motorbike |bike |public transport | walking</i></td></tr>
</table>

1. Data Introduction

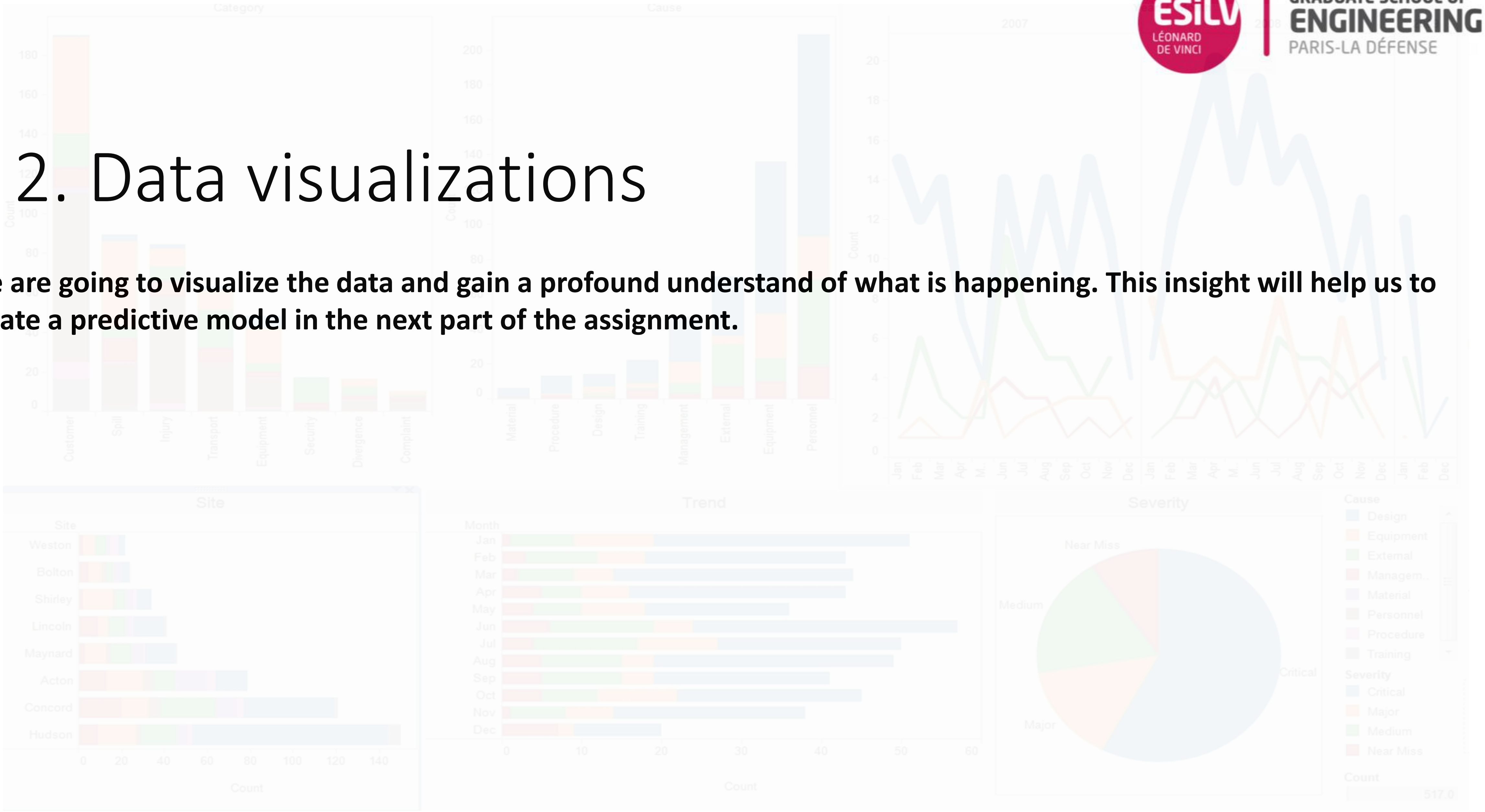
The goal :

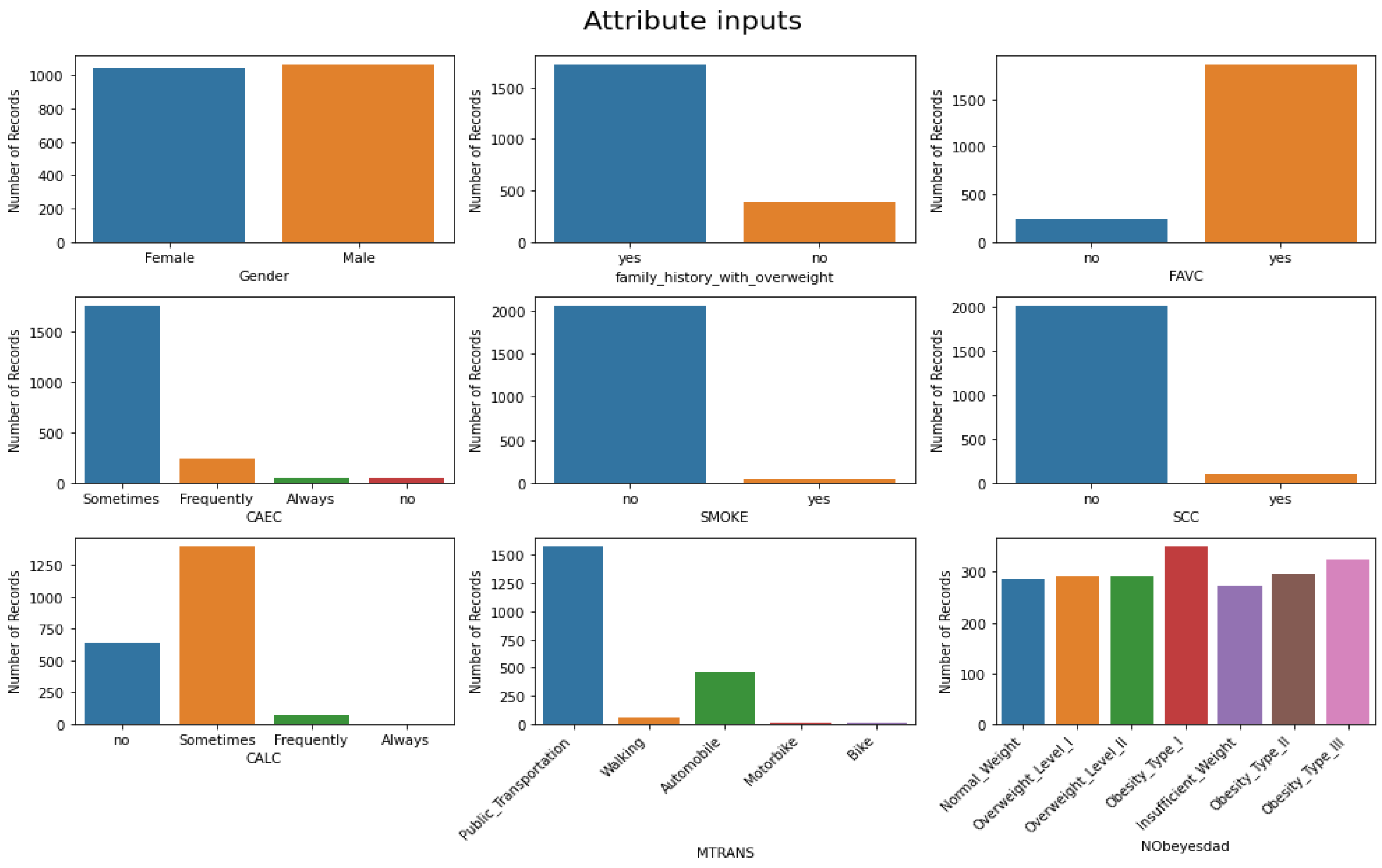
Create a prediction model able to predict in which of the obesity classification labels an individual fits.

However, since the height and weight of an individual are also attribute, the classification can easily be done using the BMI index of a person. In order to analyze the data accordingly a model will also be created while removing the two attributes able to determine the BMI.

2. Data visualizations

We are going to visualize the data and gain a profound understand of what is happening. This insight will help us to create a predictive model in the next part of the assignment.

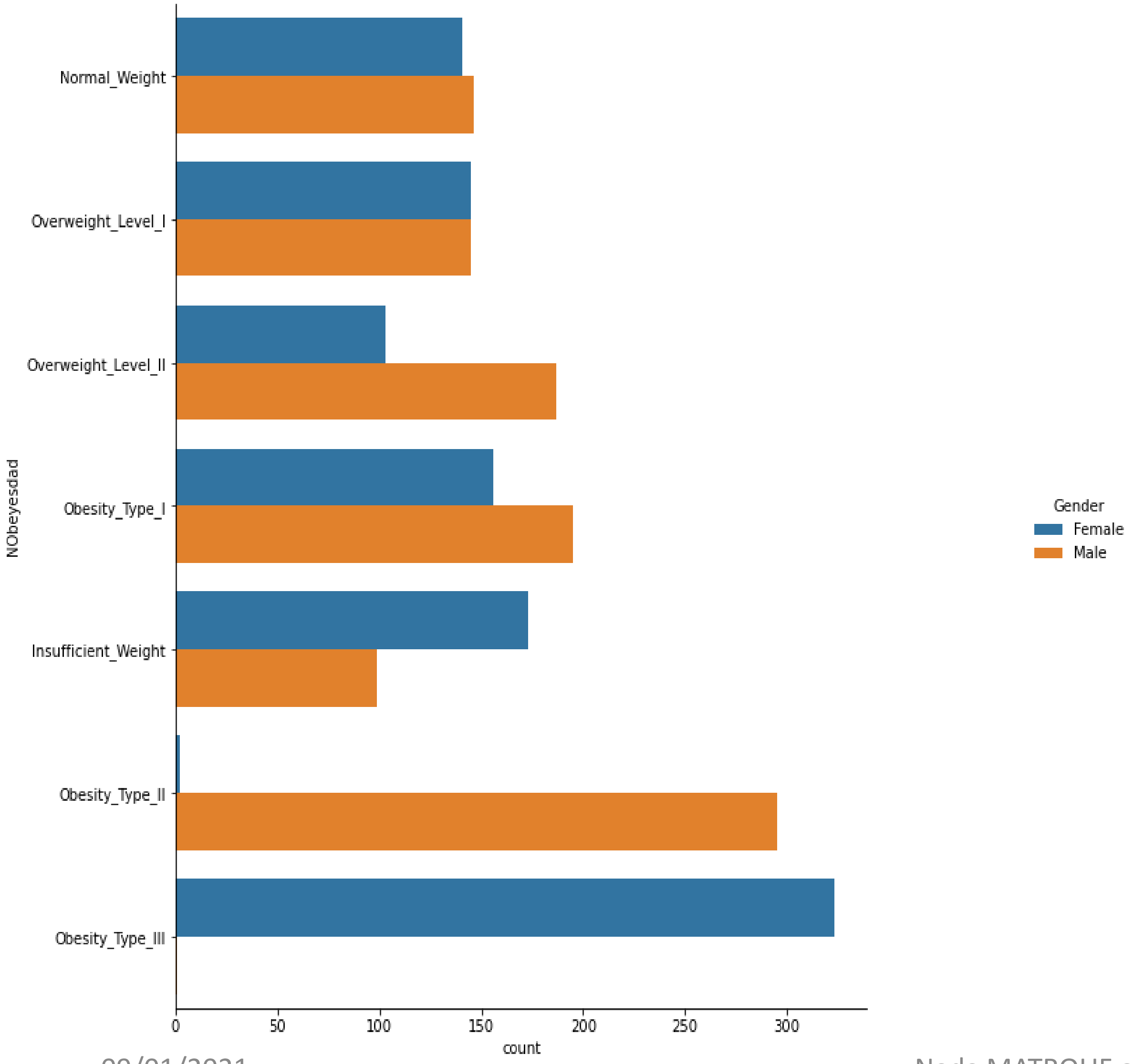




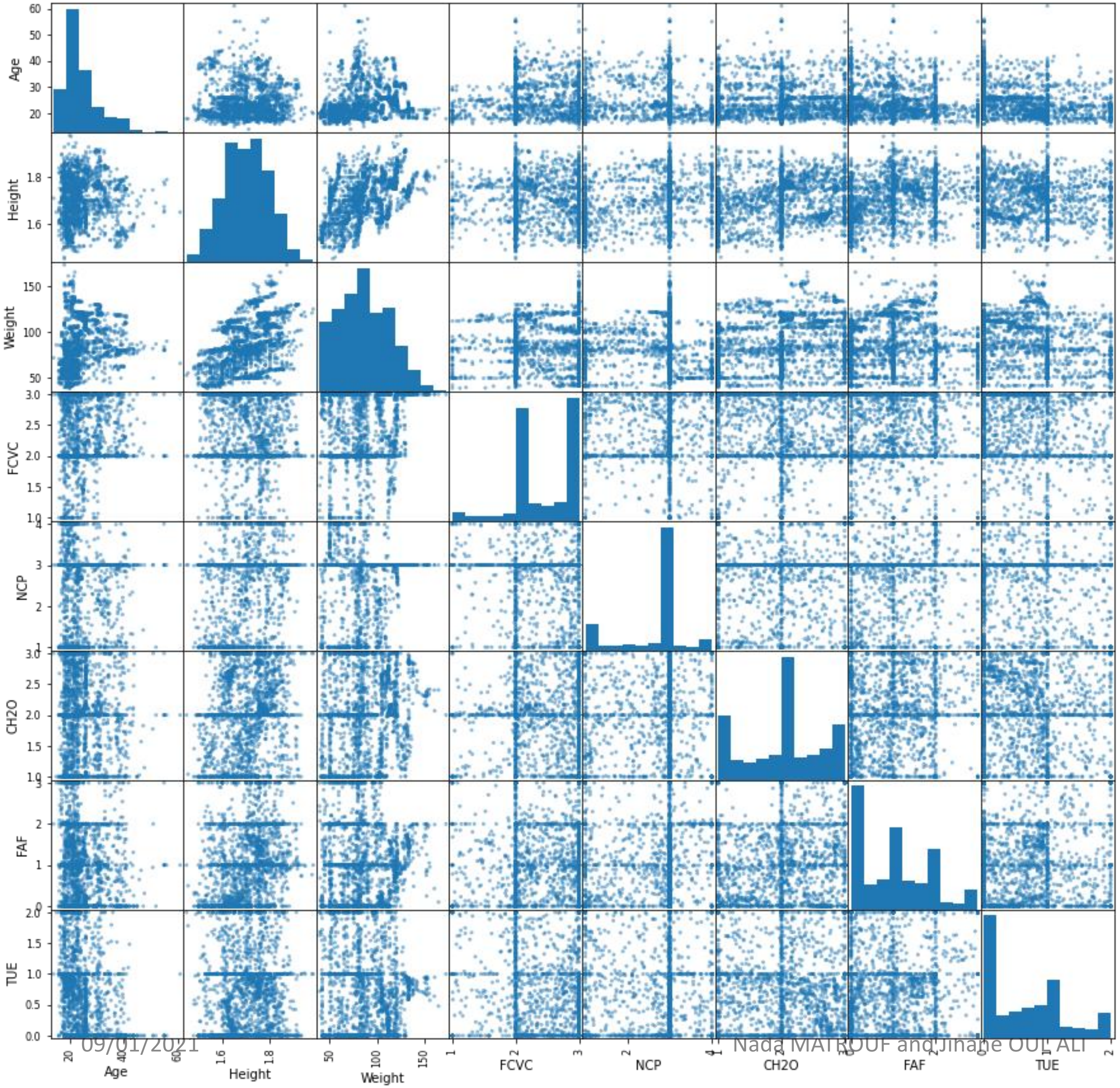
First, we will visualize the data that have various categories. This way all the different Input variables can be studied. From the histograms we can see that:

- There is an even number of female and male participant.

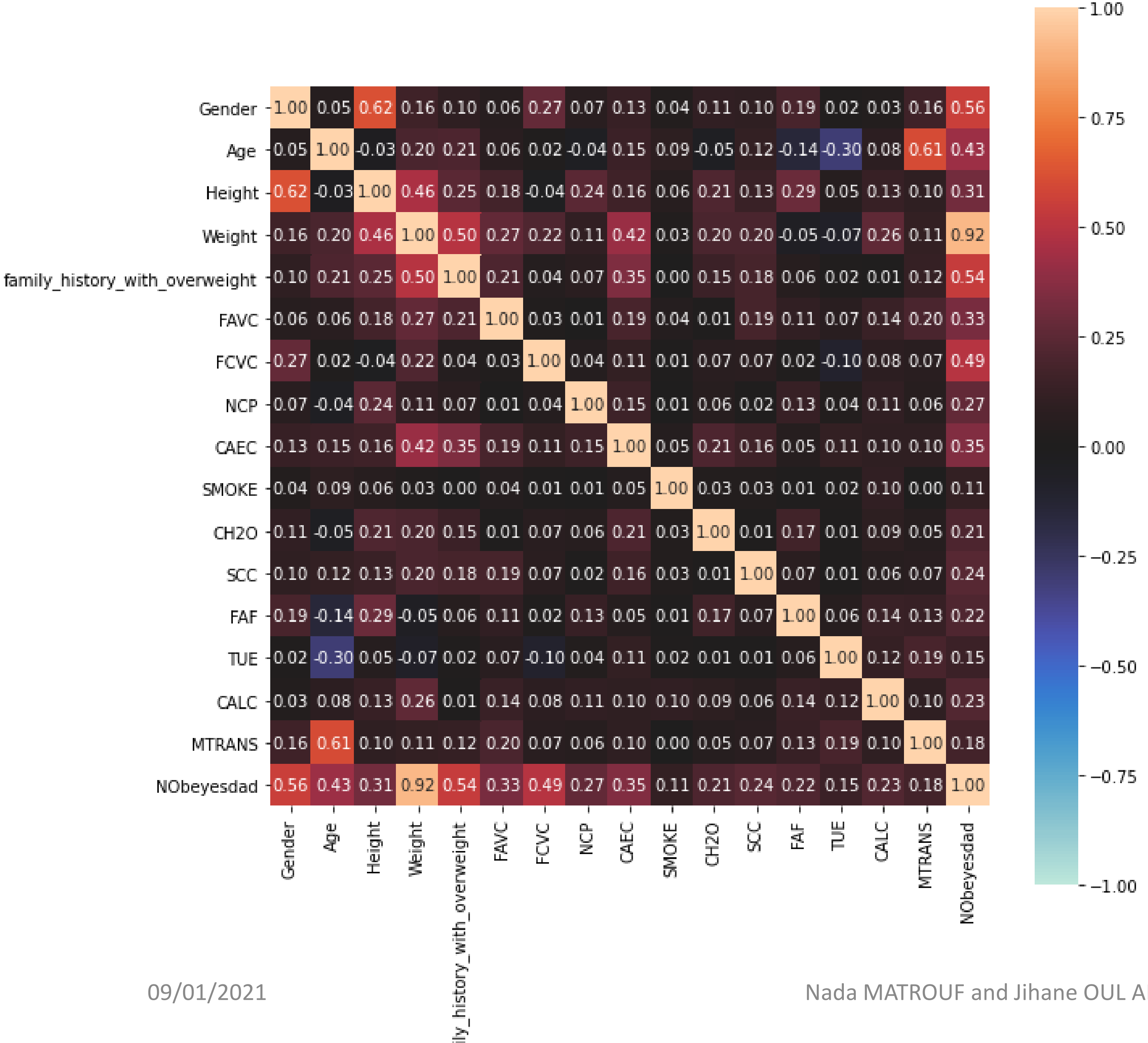
The rest of the attributes show a clear distinction between each category.



Even though the number female and male participant is the same, a clear distinction can be made in the target values for almost all of the classes except the normal weight class and the first type of overweight class. This indicates that the gender plays a very important role in assigning the labels to each individual.



The following plot shows a scatter plot of each of the 15 attributes.



	Gender	Age	Height	Weight	ily_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
Gender	1.00	0.05	0.62	0.16	0.10	0.06	0.27	0.07	0.13	0.04	0.11	0.10	0.19	0.02	0.03	0.16	0.56
Age	0.05	1.00	-0.03	0.20	0.21	0.06	0.02	-0.04	0.15	0.09	-0.05	0.12	-0.14	-0.30	0.08	0.61	0.43
Height	0.62	-0.03	1.00	0.46	0.25	0.18	-0.04	0.24	0.16	0.06	0.21	0.13	0.29	0.05	0.13	0.10	0.31
Weight	0.16	0.20	0.46	1.00	0.50	0.27	0.22	0.11	0.42	0.03	0.20	0.20	-0.05	-0.07	0.26	0.11	0.92
family_history_with_overweight	0.10	0.21	0.25	0.50	1.00	0.21	0.04	0.07	0.35	0.00	0.15	0.18	0.06	0.02	0.01	0.12	0.54
FAVC	0.06	0.06	0.18	0.27	0.21	1.00	0.03	0.01	0.19	0.04	0.01	0.19	0.11	0.07	0.14	0.20	0.33
FCVC	0.27	0.02	-0.04	0.22	0.04	0.03	1.00	0.04	0.11	0.01	0.07	0.07	0.02	-0.10	0.08	0.07	0.49
NCP	0.07	-0.04	0.24	0.11	0.07	0.01	0.04	1.00	0.15	0.01	0.06	0.02	0.13	0.04	0.11	0.06	0.27
CAEC	0.13	0.15	0.16	0.42	0.35	0.19	0.11	0.15	1.00	0.05	0.21	0.16	0.05	0.11	0.10	0.10	0.35
SMOKE	0.04	0.09	0.06	0.03	0.00	0.04	0.01	0.01	0.05	1.00	0.03	0.03	0.01	0.02	0.10	0.00	0.11
CH2O	0.11	-0.05	0.21	0.20	0.15	0.01	0.07	0.06	0.21	0.03	1.00	0.01	0.17	0.01	0.09	0.05	0.21
SCC	0.10	0.12	0.13	0.20	0.18	0.19	0.07	0.02	0.16	0.03	0.01	1.00	0.07	0.01	0.06	0.07	0.24
FAF	0.19	-0.14	0.29	-0.05	0.06	0.11	0.02	0.13	0.05	0.01	0.17	0.07	1.00	0.06	0.14	0.13	0.22
TUE	0.02	-0.30	0.05	-0.07	0.02	0.07	-0.10	0.04	0.11	0.02	0.01	0.01	0.06	1.00	0.12	0.19	0.15
CALC	0.03	0.08	0.13	0.26	0.01	0.14	0.08	0.11	0.10	0.10	0.09	0.06	0.14	0.12	1.00	0.10	0.23
MTRANS	0.16	0.61	0.10	0.11	0.12	0.20	0.07	0.06	0.10	0.00	0.05	0.07	0.13	0.19	0.10	1.00	0.18
NObeyesdad	0.56	0.43	0.31	0.92	0.54	0.33	0.49	0.27	0.35	0.11	0.21	0.24	0.22	0.15	0.23	0.18	1.00

A heat map is also created. The heatmap shows us the correlation between each attribute and therefore its importance in the prediction model.

As previously explained, the correlation between the target attribute and weight seems very high. Which agrees with our previously made hypothesis that the results heavenly rely on the BMI.

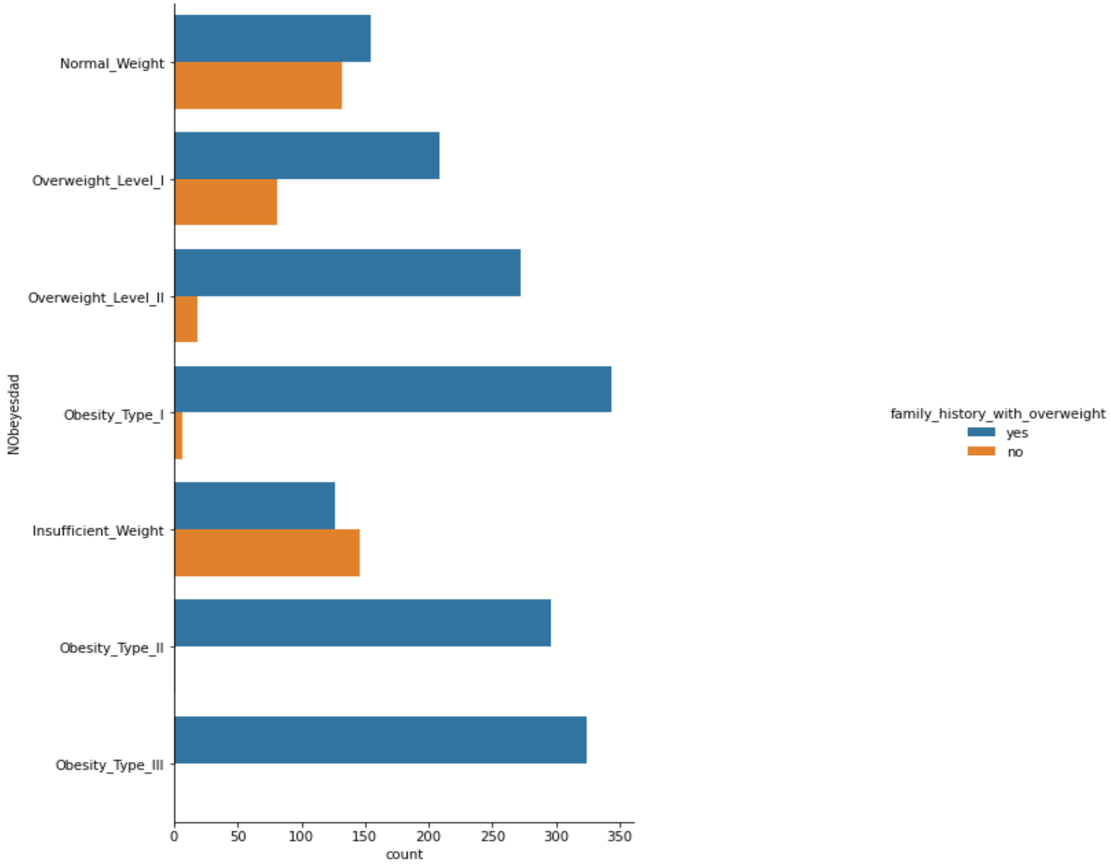
However, the heat map also shows a high correlation between the Gender and the target attribute NObeyesdad. The next important attributes are :

Family history with, overweight,FCVC,AGE,CAEC,FAVC in that order.

Some of these attributes will be looked at closer during the visualization phase to gain more insight.

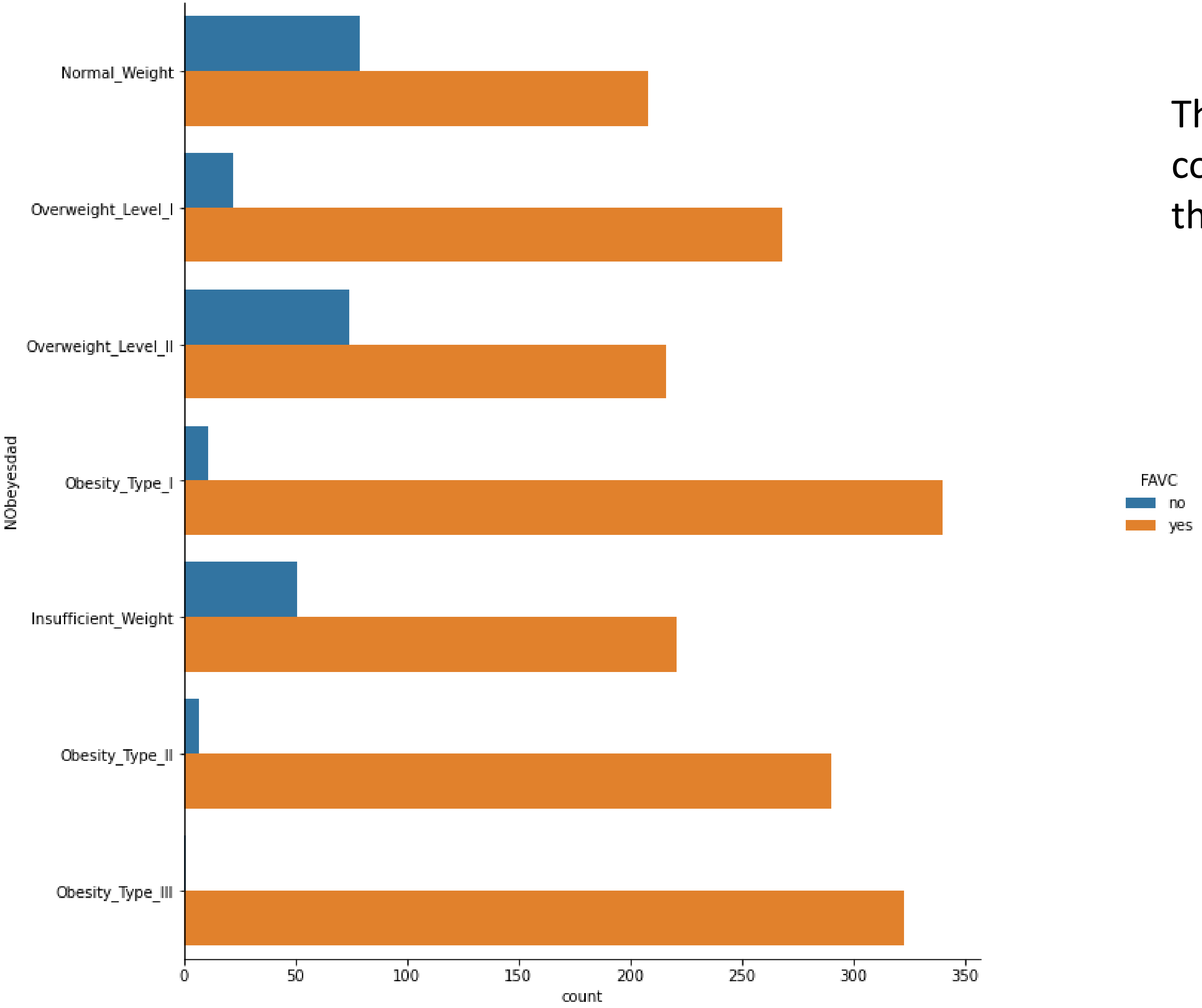
Family history with overweight

First,we will start with the attribute showcasing the family history with overweight. The input values for this attribute is Boolean.



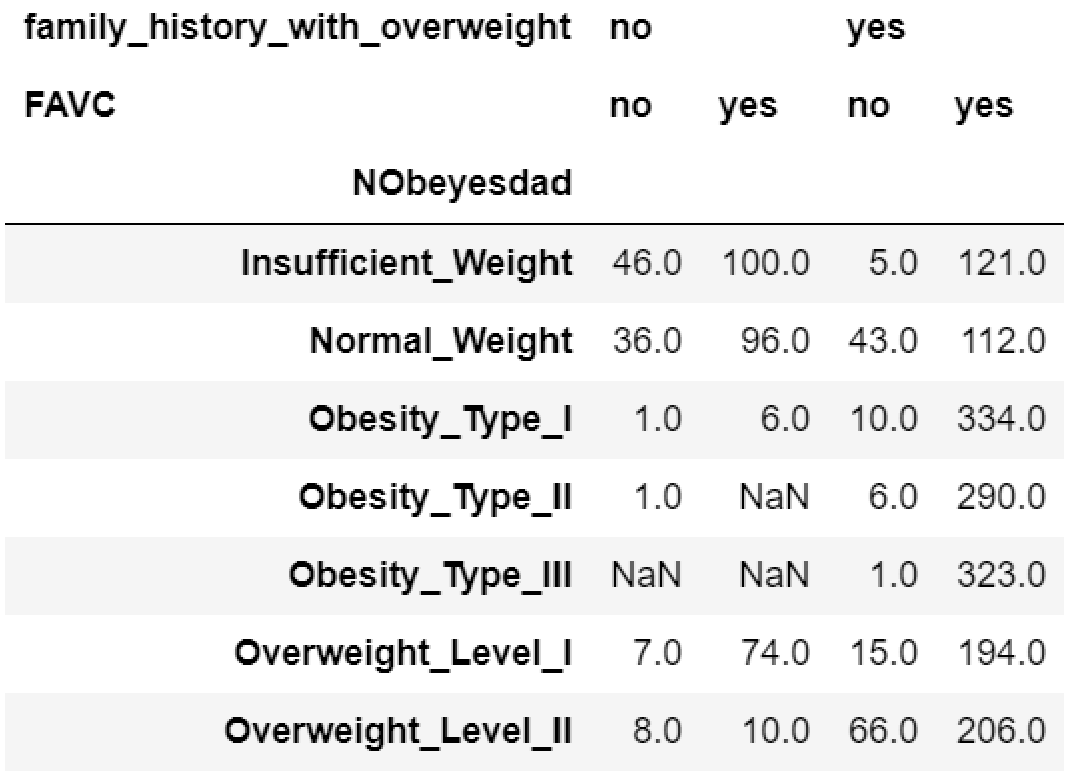
We can clearly see that although for all the target except obesity type 2 and obesity type 3, there are some individuals suffering even though no history of overweight family members is detected. However, it can't be denied that the influence of family history is huge on all types of obesity and overweight type 1 and 2. This agrees with the many studies performed previously on the link of overweight family history and severe onset of obeisant. Below links of some of the studies performed:

Frequency consumption of high caloric food



The next step is to determine how the frequent consumption of food high in calories influences the various types of target labels:

Here we can see that most of the people answered with yes which mean the high amount of calories does not necessarily influence the output. However, it can be noticed that the people with a normal weight do have a tendency to eat less high caloric meals on a daily basis.

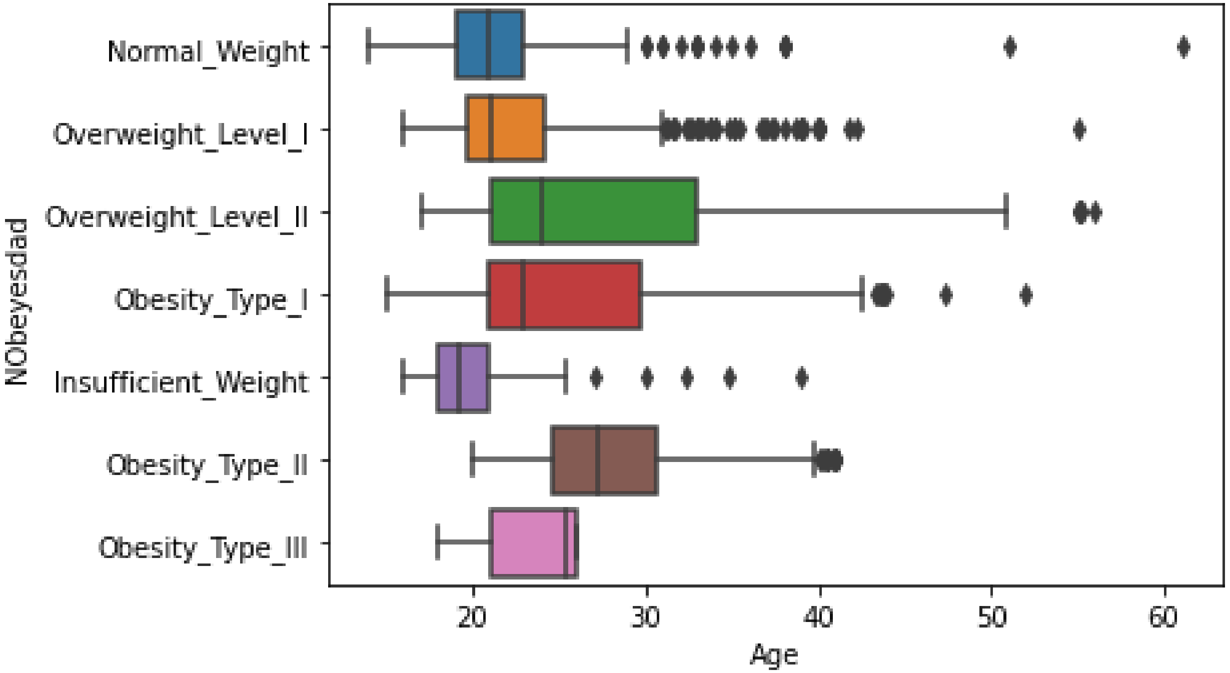


family_history_with_overweight	no		yes	
FAVC	no	yes	no	yes
NObeyesdad				
Insufficient_Weight	46.0	100.0	5.0	121.0
Normal_Weight	36.0	96.0	43.0	112.0
Obesity_Type_I	1.0	6.0	10.0	334.0
Obesity_Type_II	1.0	NaN	6.0	290.0
Obesity_Type_III	NaN	NaN	1.0	323.0
Overweight_Level_I	7.0	74.0	15.0	194.0
Overweight_Level_II	8.0	10.0	66.0	206.0

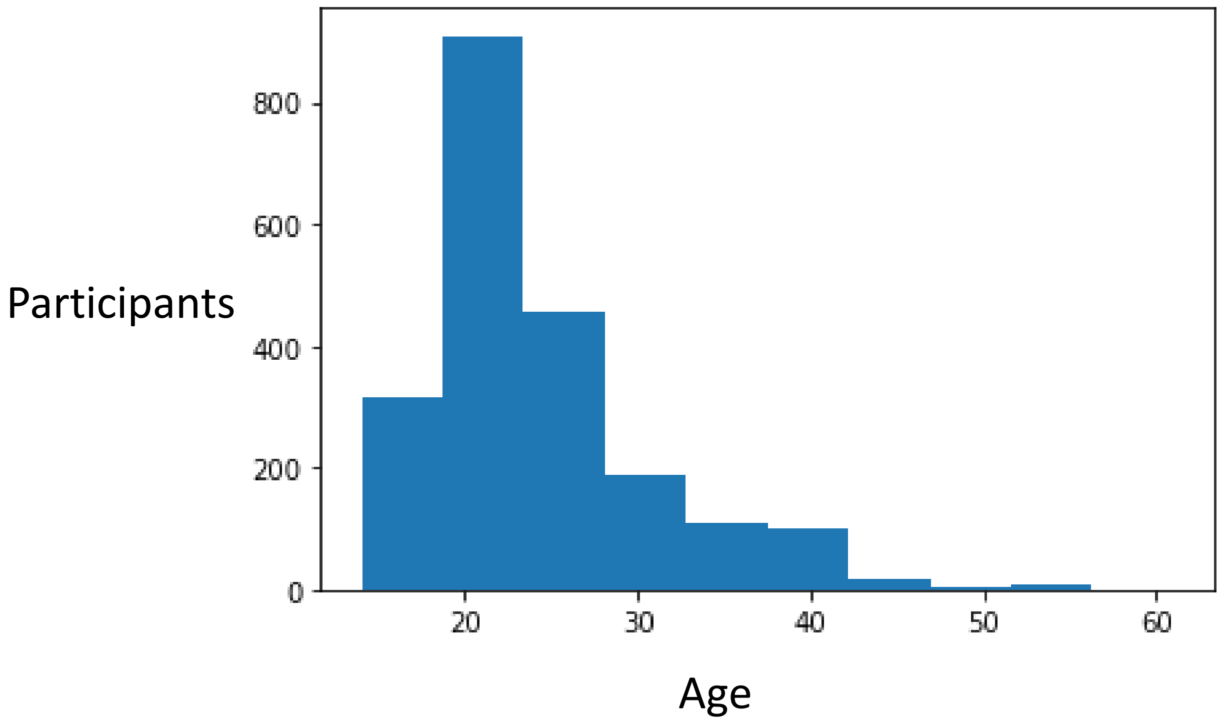
Here we can see that most of the people that have a family members with history of obesity also tend to consume high calorie meals on a daily bases. The majority of these people have type 1 obesity closely followed by type 3 obesity. This is a very important remark since it tends to also show the link between having a family member that suffered or is suffering from obesity and the tendency to eat more.

Age

Next we will look at the age. The following boxplots will show the average age for each target class.



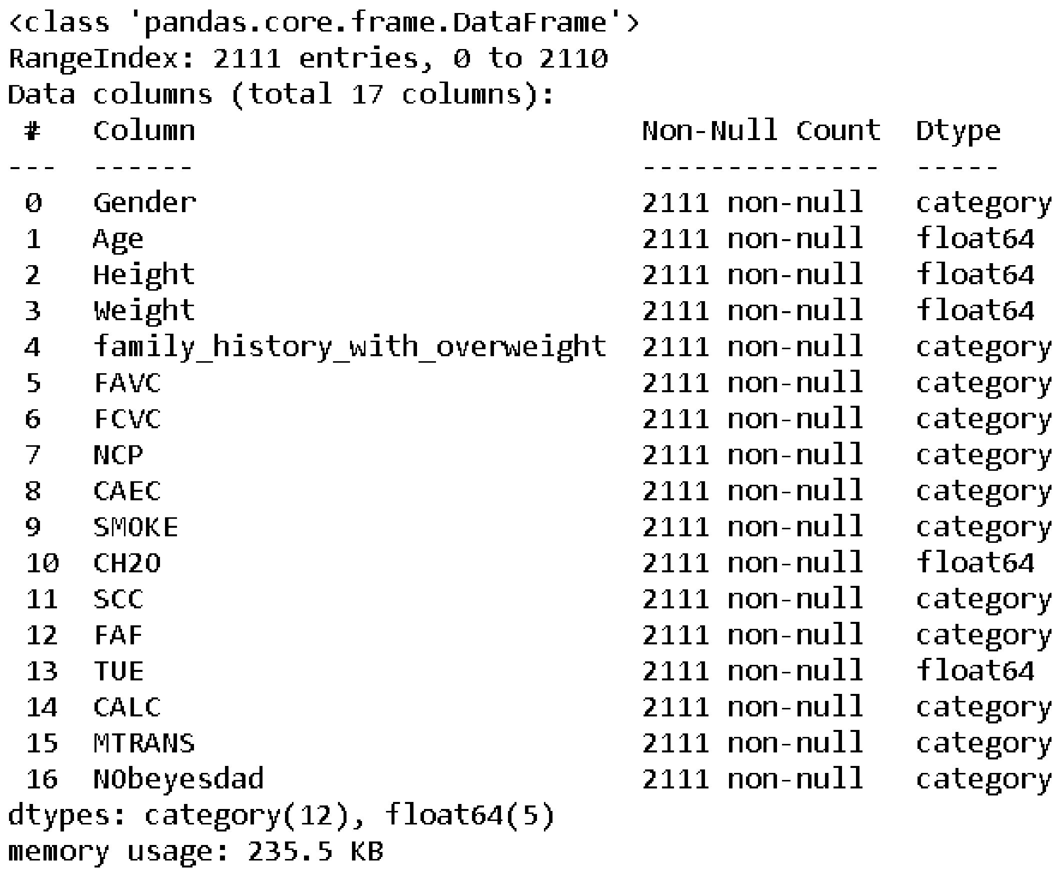
We can see that most of the targets have an average age between 20 and 30. This sounds counter intuitive since most of the people at that age are at their fittest. Therefore, we will plot a graph showcasing the age of the individuals.



The plot shows that the majority of the participants is between 20 and 30. This explains our previously found conclusions. However, the boxplots are not completely useless. The boxplots showcasing an average that leans toward the 30 indicates a very strong influence for the people that are above 30. Since this group of participants is small the numbers must be high. These include obesity type 2 and type 3.

3. Data Preparation

As explained previously, we will create two models one with the height and weight and one without. In addition, the different strings in the attributes will be transformed to categories. (Other encoding processes are possible with different libraries however this seemed to be the most straightforward)



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2111 non-null   category
 1   Age                             2111 non-null   float64
 2   Height                          2111 non-null   float64
 3   Weight                          2111 non-null   float64
 4   family_history_with_overweight  2111 non-null   category
 5   FAVC                            2111 non-null   category
 6   FCVC                            2111 non-null   category
 7   NCP                             2111 non-null   category
 8   CAEC                            2111 non-null   category
 9   SMOKE                           2111 non-null   category
 10  CH2O                            2111 non-null   float64
 11  SCC                             2111 non-null   category
 12  FAF                             2111 non-null   category
 13  TUE                             2111 non-null   float64
 14  CALC                            2111 non-null   category
 15  MTRANS                          2111 non-null   category
 16  NObeyesdad                      2111 non-null   category
dtypes: category(12), float64(5)
memory usage: 235.5 KB
```

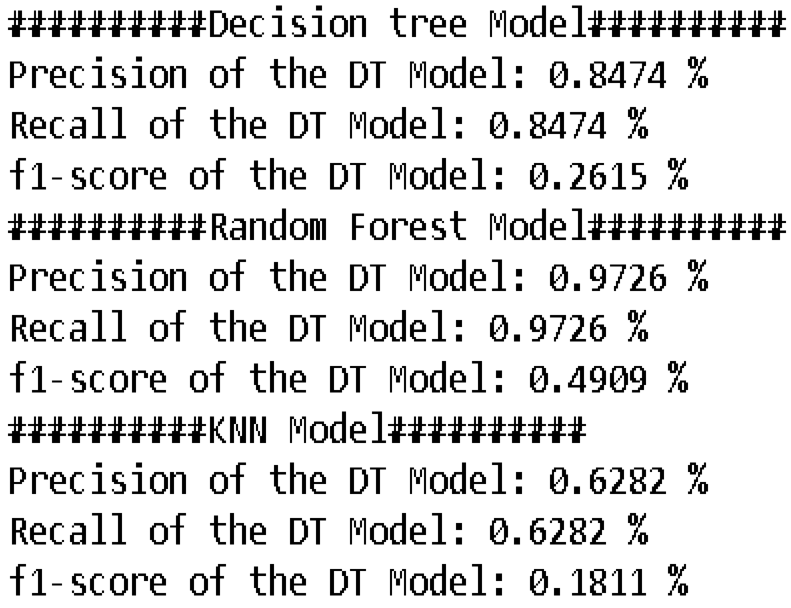
We will also use encoding for the CAEC and CALC perimeters since the degree here is important.

4. Modelisation and Validation

In order to build a prediction model many machine learning algorithms can be applied. For this specific application the goal is to create a prediction model able to classify the individual based on the given attribute. This leaves us with the many options for classification algorithms. For this approach three types will be considered:

- **Decision tree model** : In this model the algorithm creates many subtrees depending on the attribute input.
- **Random Forest model** : This is an extension of the decision tree model. The Random Forest Algorithm combines the output of multiple (randomly created) Decision Trees to generate the final output.
- **KNN** : KNN works by finding the distance between data points by selecting a number K of groups and classifying them by distance.

Dataset without the height and the weight

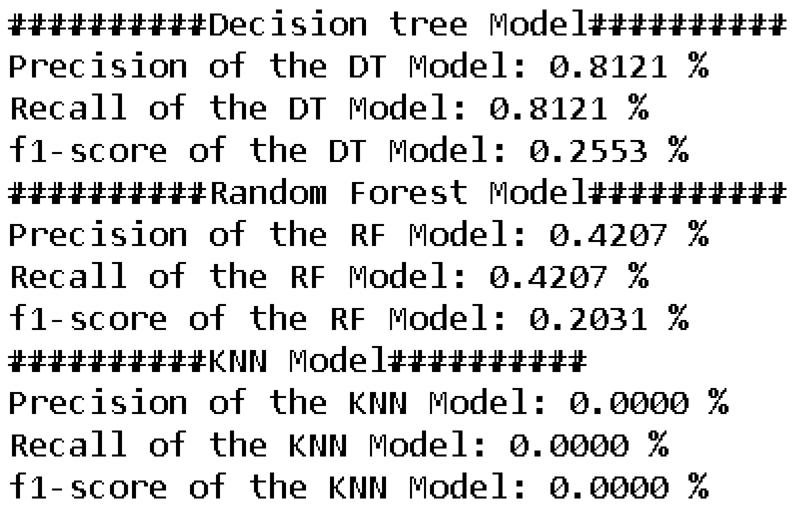


```
##########Decision tree Model##########
Precision of the DT Model: 0.8474 %
Recall of the DT Model: 0.8474 %
f1-score of the DT Model: 0.2615 %
##########Random Forest Model##########
Precision of the DT Model: 0.9726 %
Recall of the DT Model: 0.9726 %
f1-score of the DT Model: 0.4909 %
##########KNN Model##########
Precision of the DT Model: 0.6282 %
Recall of the DT Model: 0.6282 %
f1-score of the DT Model: 0.1811 %
```

From these results, we can conclude that the Random Forest Model proved to be the best model for our data set with an accuracy of 97%.

Dataset with the height and the weight

Next we will look at the dataset while including the height and the weight.



```
##########Decision tree Model##########
Precision of the DT Model: 0.8121 %
Recall of the DT Model: 0.8121 %
f1-score of the DT Model: 0.2553 %
##########Random Forest Model##########
Precision of the RF Model: 0.4207 %
Recall of the RF Model: 0.4207 %
f1-score of the RF Model: 0.2031 %
##########KNN Model##########
Precision of the KNN Model: 0.0000 %
Recall of the KNN Model: 0.0000 %
f1-score of the KNN Model: 0.0000 %
```

5. Conclusion

Although for the last models the height and weight were added, the predictions were not more accurate. The Random Forest Model scored the highest with an accuracy of 93%. However, the Random Forest Model accuracy made without the two attributes : height and weight is still higher (97%). Therefore, the Random Forest Model will be chosen for the rest of this assignment.

References

Special thanks to

Fabio Mendoza Palechor and Alexis de la Hoz Manotas

For making the dataset available