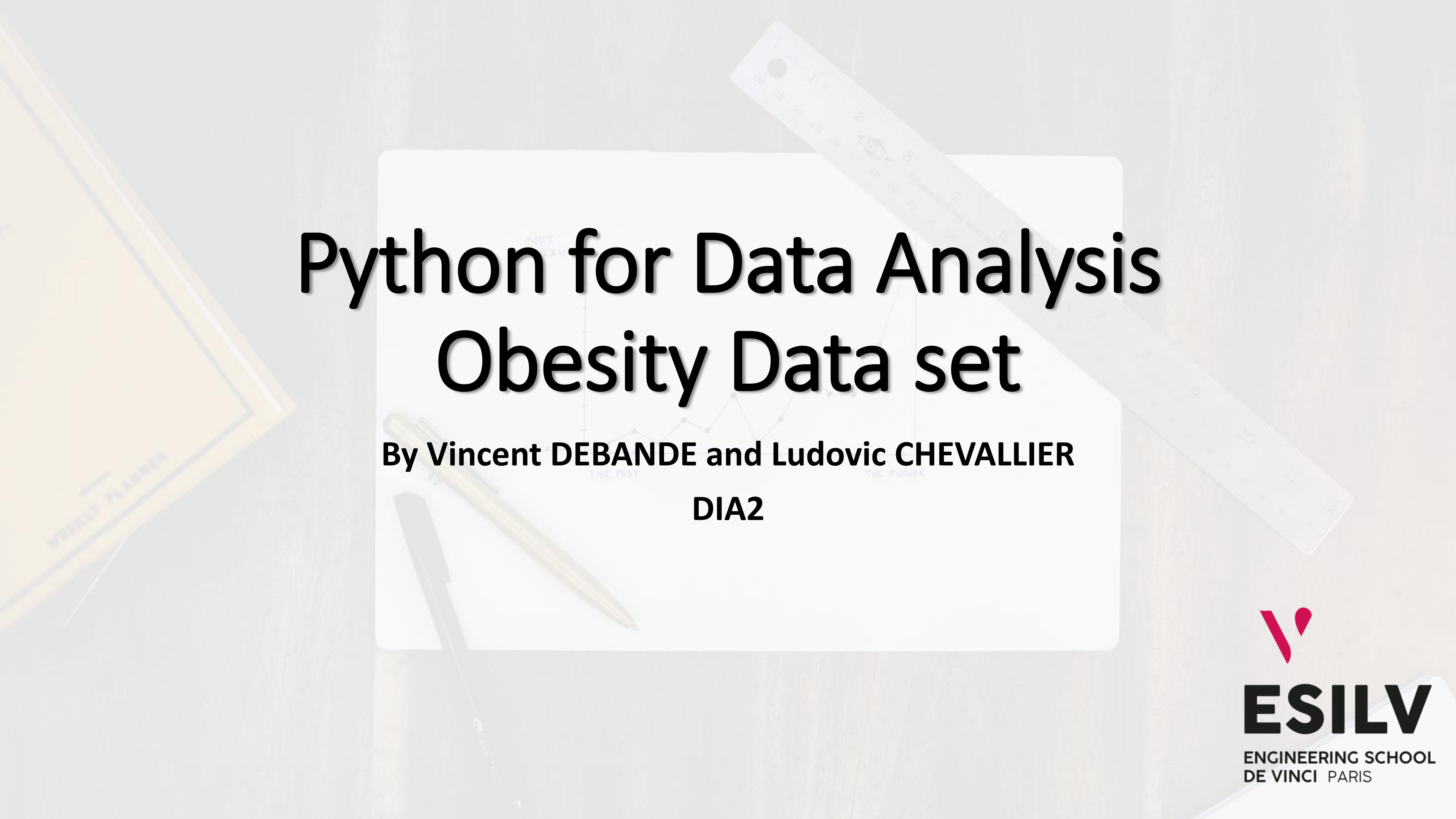


# Python for Data Analysis
# Obesity Data set

**By Vincent DEBANDE and Ludovic CHEVALLIER**

**DIA2**

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

# 1 - The Study *(Obesity, 28-07-2019)*

## Context:

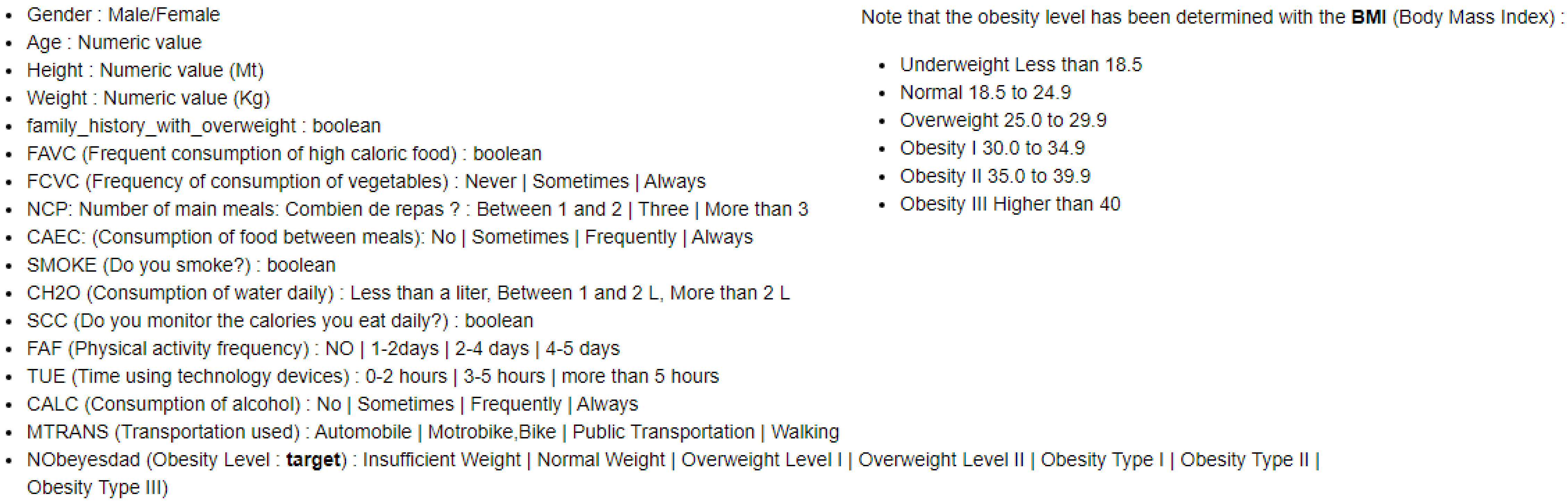
- Obesity as doubled since 1980.
- The objective is to determine if one person suffers from obesity based on a lot of factors (not always related to eating habits) in South America.

## Data set:

- 77% of the data was generated synthetically using the Weka tool and the SMOTE filter, 23% of the data was collected directly from users through a web platform.
- The data contains 17 attributes and 2111 records

# 2 - Analysis of the attributes

- The very first step was to understand the data :

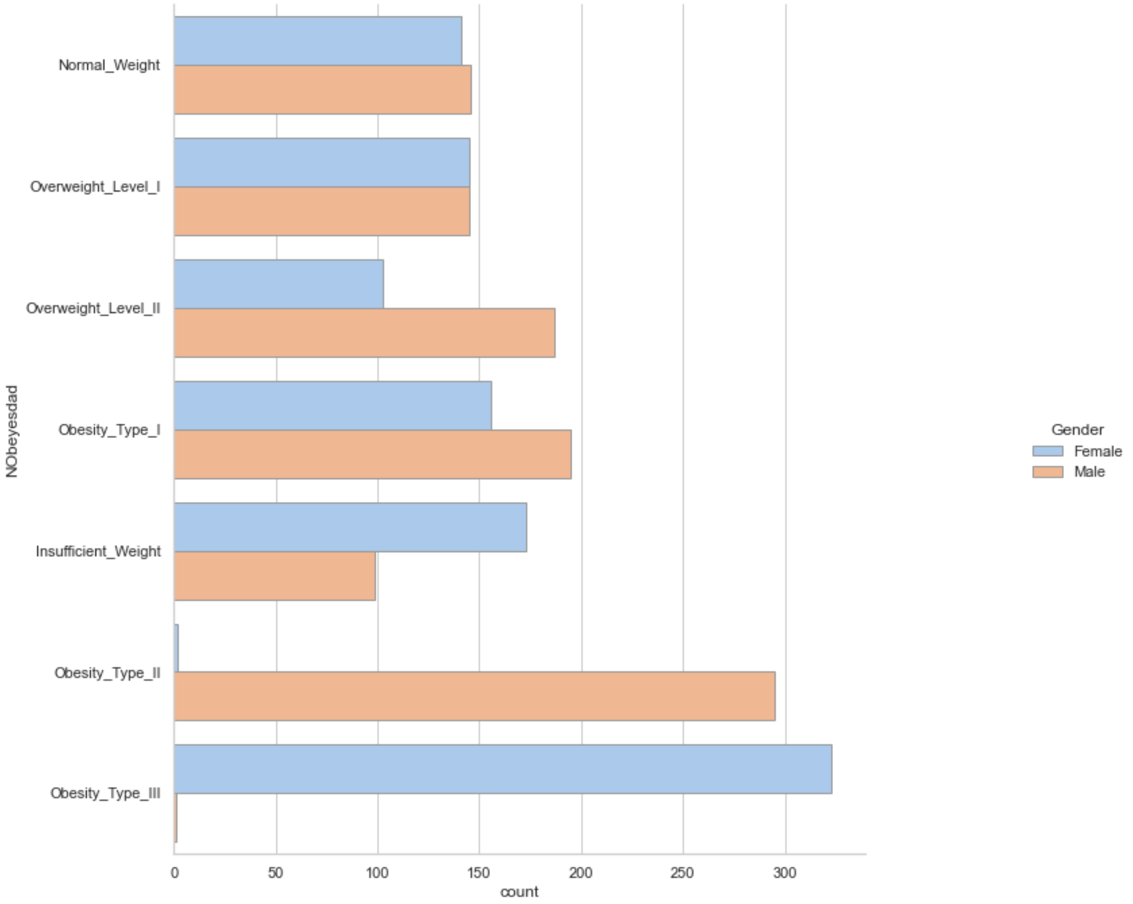


- Gender : Male/Female
- Age : Numeric value
- Height : Numeric value (Mt)
- Weight : Numeric value (Kg)
- family_history_with_overweight : boolean
- FAVC (Frequent consumption of high caloric food) : boolean
- FCVC (Frequency of consumption of vegetables) : Never | Sometimes | Always
- NCP: Number of main meals: Combien de repas ? : Between 1 and 2 | Three | More than 3
- CAEC: (Consumption of food between meals): No | Sometimes | Frequently | Always
- SMOKE (Do you smoke?) : boolean
- CH2O (Consumption of water daily) : Less than a liter, Between 1 and 2 L, More than 2 L
- SCC (Do you monitor the calories you eat daily?) : boolean
- FAF (Physical activity frequency) : NO | 1-2days | 2-4 days | 4-5 days
- TUE (Time using technology devices) : 0-2 hours | 3-5 hours | more than 5 hours
- CALC (Consumption of alcohol) : No | Sometimes | Frequently | Always
- MTRANS (Transportation used) : Automobile | Motrobike,Bike | Public Transportation | Walking
- NObeyesdad (Obesity Level : **target**) : Insufficient Weight | Normal Weight | Overweight Level I | Overweight Level II | Obesity Type I | Obesity Type II | Obesity Type III)

Note that the obesity level has been determined with the **BMI** (Body Mass Index) :

- Underweight Less than 18.5
- Normal 18.5 to 24.9
- Overweight 25.0 to 29.9
- Obesity I 30.0 to 34.9
- Obesity II 35.0 to 39.9
- Obesity III Higher than 40

# 2.1 – Gender anomaly



We can see that the class Obesity_Type_II is composed only of Male and Type_III only of females.

After computation of Gower's Distance (mean of 0.20) for each individual of Type_II to all of Type_III we saw that they were similar apart from : Height et Weight.

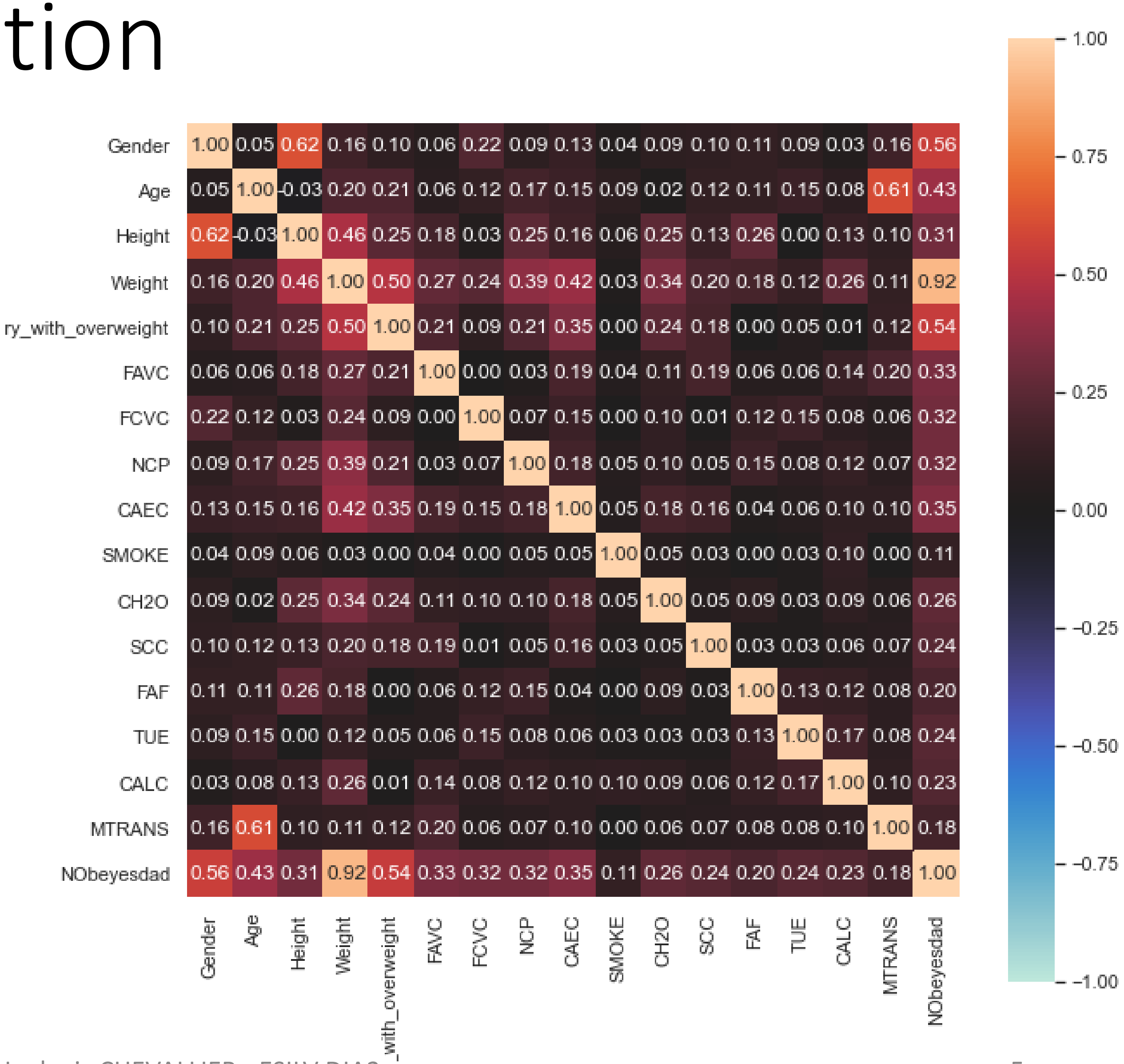
Female are shorter and a little bit heavier, so their BMI is higher. But it's not normal that there is not the opposite gender in those categories.

Since the BMI is unisex, we decided to move on.

# 2.2 – Features correlation

- Using Theil's U and Cramer's V, we were able to see the most important features. We analyzed them deeply in our Notebook.

- We'll remove Height and Weight from our data set because with those features, there is no need for ML.



| | Gender | Age | Height | Weight | _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 | 0.22 | 0.09 | 0.13 | 0.04 | 0.09 | 0.10 | 0.11 | 0.09 | 0.03 | 0.16 | 0.56 |
| Age | 0.05 | 1.00 | -0.03 | 0.20 | 0.21 | 0.06 | 0.12 | 0.17 | 0.15 | 0.09 | 0.02 | 0.12 | 0.11 | 0.15 | 0.08 | 0.61 | 0.43 |
| Height | 0.62 | -0.03 | 1.00 | 0.46 | 0.25 | 0.18 | 0.03 | 0.25 | 0.16 | 0.06 | 0.25 | 0.13 | 0.26 | 0.00 | 0.13 | 0.10 | 0.31 |
| Weight | 0.16 | 0.20 | 0.46 | 1.00 | 0.50 | 0.27 | 0.24 | 0.39 | 0.42 | 0.03 | 0.34 | 0.20 | 0.18 | 0.12 | 0.26 | 0.11 | 0.92 |
| ry_with_overweight | 0.10 | 0.21 | 0.25 | 0.50 | 1.00 | 0.21 | 0.09 | 0.21 | 0.35 | 0.00 | 0.24 | 0.18 | 0.00 | 0.05 | 0.01 | 0.12 | 0.54 |
| FAVC | 0.06 | 0.06 | 0.18 | 0.27 | 0.21 | 1.00 | 0.00 | 0.03 | 0.19 | 0.04 | 0.11 | 0.19 | 0.06 | 0.06 | 0.14 | 0.20 | 0.33 |
| FCVC | 0.22 | 0.12 | 0.03 | 0.24 | 0.09 | 0.00 | 1.00 | 0.07 | 0.15 | 0.00 | 0.10 | 0.01 | 0.12 | 0.15 | 0.08 | 0.06 | 0.32 |
| NCP | 0.09 | 0.17 | 0.25 | 0.39 | 0.21 | 0.03 | 0.07 | 1.00 | 0.18 | 0.05 | 0.10 | 0.05 | 0.15 | 0.08 | 0.12 | 0.07 | 0.32 |
| CAEC | 0.13 | 0.15 | 0.16 | 0.42 | 0.35 | 0.19 | 0.15 | 0.18 | 1.00 | 0.05 | 0.18 | 0.16 | 0.04 | 0.06 | 0.10 | 0.10 | 0.35 |
| SMOKE | 0.04 | 0.09 | 0.06 | 0.03 | 0.00 | 0.04 | 0.00 | 0.05 | 0.05 | 1.00 | 0.05 | 0.03 | 0.00 | 0.03 | 0.10 | 0.00 | 0.11 |
| CH2O | 0.09 | 0.02 | 0.25 | 0.34 | 0.24 | 0.11 | 0.10 | 0.10 | 0.18 | 0.05 | 1.00 | 0.05 | 0.09 | 0.03 | 0.09 | 0.06 | 0.26 |
| SCC | 0.10 | 0.12 | 0.13 | 0.20 | 0.18 | 0.19 | 0.01 | 0.05 | 0.16 | 0.03 | 0.05 | 1.00 | 0.03 | 0.03 | 0.06 | 0.07 | 0.24 |
| FAF | 0.11 | 0.11 | 0.26 | 0.18 | 0.00 | 0.06 | 0.12 | 0.15 | 0.04 | 0.00 | 0.09 | 0.03 | 1.00 | 0.13 | 0.12 | 0.08 | 0.20 |
| TUE | 0.09 | 0.15 | 0.00 | 0.12 | 0.05 | 0.06 | 0.15 | 0.08 | 0.06 | 0.03 | 0.03 | 0.03 | 0.13 | 1.00 | 0.17 | 0.08 | 0.24 |
| CALC | 0.03 | 0.08 | 0.13 | 0.26 | 0.01 | 0.14 | 0.08 | 0.12 | 0.10 | 0.10 | 0.09 | 0.06 | 0.12 | 0.17 | 1.00 | 0.10 | 0.23 |
| MTRANS | 0.16 | 0.61 | 0.10 | 0.11 | 0.12 | 0.20 | 0.06 | 0.07 | 0.10 | 0.00 | 0.06 | 0.07 | 0.08 | 0.08 | 0.10 | 1.00 | 0.18 |
| NObeyesdad | 0.56 | 0.43 | 0.31 | 0.92 | 0.54 | 0.33 | 0.32 | 0.32 | 0.35 | 0.11 | 0.26 | 0.24 | 0.20 | 0.24 | 0.23 | 0.18 | 1.00 |

# 3 – Data Preparation (Encoding)

Since our data is mainly (apart from Age) categorical, we needed to distinguish **Nominal** (with no order, such as colors) features from **Ordinal** (with order such as frequency or level of education) ones.

| Nominal |
|---|
| Gender<br>FAVC (Frequent consumption of high caloric food)<br>family_history_with_overweight<br>SMOKE<br>SCC (Do you monitor the calories you eat daily?)<br>MTRANS (Transportation used) |

# 3.1 – Encoding

- We encoded Ordinal variables using Label Encoding (numbers for each values)

- Nominal variables were encoded using One Hot Encoding (multiple columns for each values. “Dummy values”)

# 4 – Machine Learning

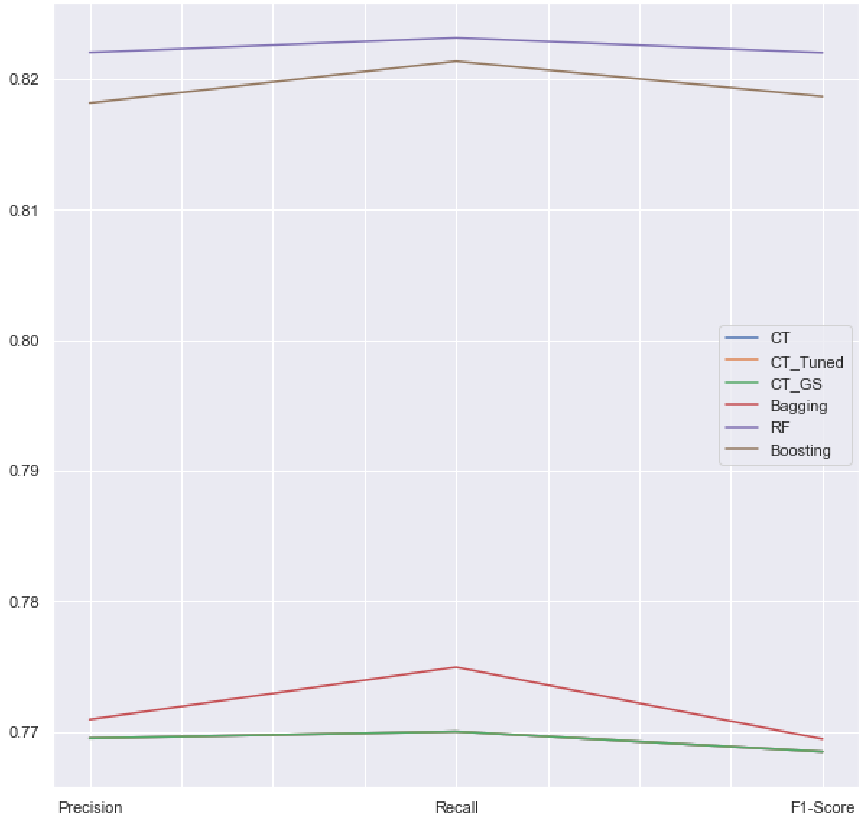
Since we have a classification problem for multi class, we decided to try LDA (Linear Discriminant Analysis) and Trees (Classification Tree, Bagging, Random Forest and Boosting)

We also tuned and performed Grid Search some models in order to have better performance.

| Model | Accuracy Test | Accuracy Train |
|---|---|---|
| LDA | 57.68% | 57.52% |
| Classification Tree \| Tuned \| GS | 77.54% \| 73.52% \| 73.52% | 96.8% \| 91.94% \| 92% |
| Bagging | 78.01% | 95.56% |
| Random Forest | 82.51% | 96.8% |
| Boosting \| Tuned \| GS | 82.27% \| 51.56% \| 82.51% | 96.8% \| 96.8% \| 96.8% |

# 4.1 Model comparison

For each model and each class, we have : Precision, Recall, F1-Score. We computed the mean for each model and did a plot to compare them.



We finally chose the untuned Random Forest model because of its performances and its low runtime (compared to Boosting).

# Credits – Thank for data utilization

Thank to :

- Palechor, F. M., & de la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in Brief, 104344.