Python Project : ObesityDataSet_raw_and_data_sinthetic

BERTHAULT Quentin - BITAR Aref

The Dataset

- The dataset is composed of 17 columns and 2111 rows
- On the 17 columns 8 are floats and 9 are objects
- The dataset is about the habits or the physical characteristics of 2111 persons and the last column is about their obesity type which is what we are going to focus on.

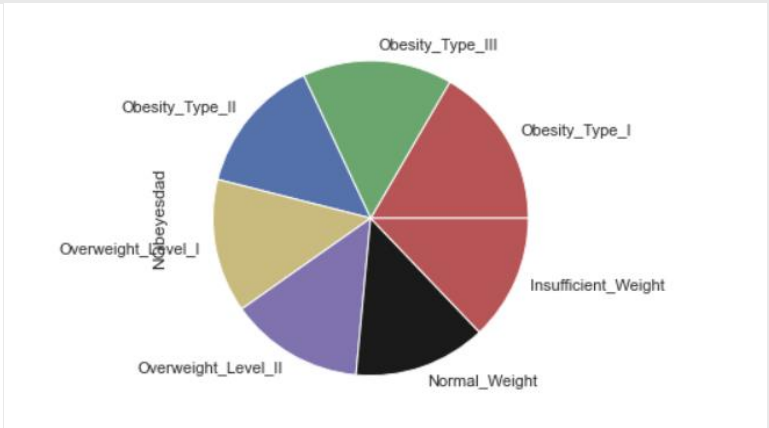
```
print(data.shape)
```

```
(2111, 17)
```

```
object     9
float64    8
```

Type of obesity (7 types) :

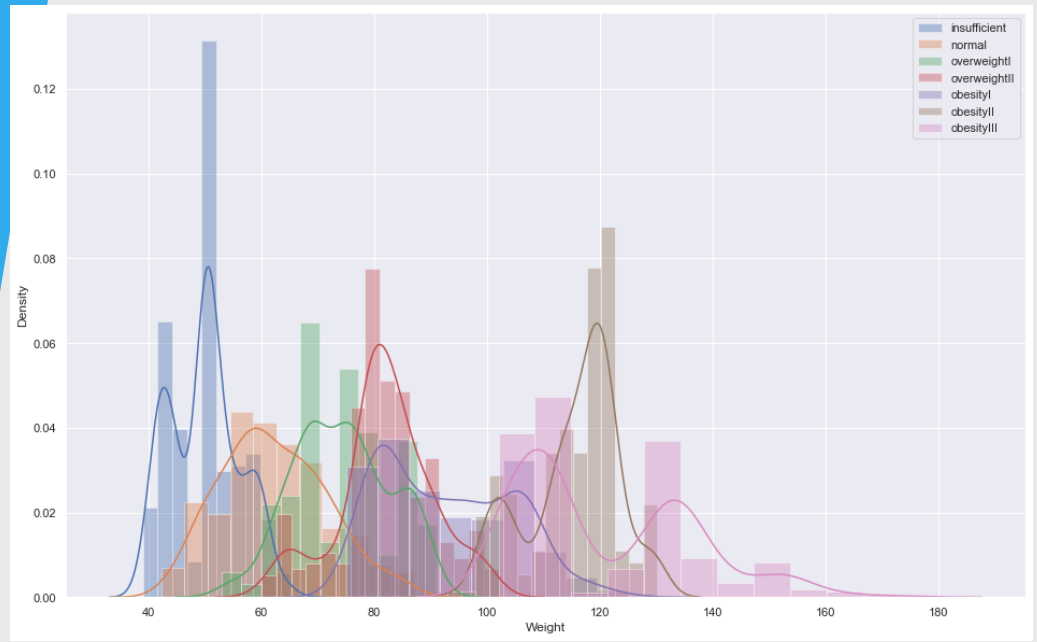
```
Valeurs numériques :

Obesity_Type_I          351
Obesity_Type_III        324
Obesity_Type_II         297
Overweight_Level_I      290
Overweight_Level_II     290
Normal_Weight           287
Insufficient_Weight     272
```

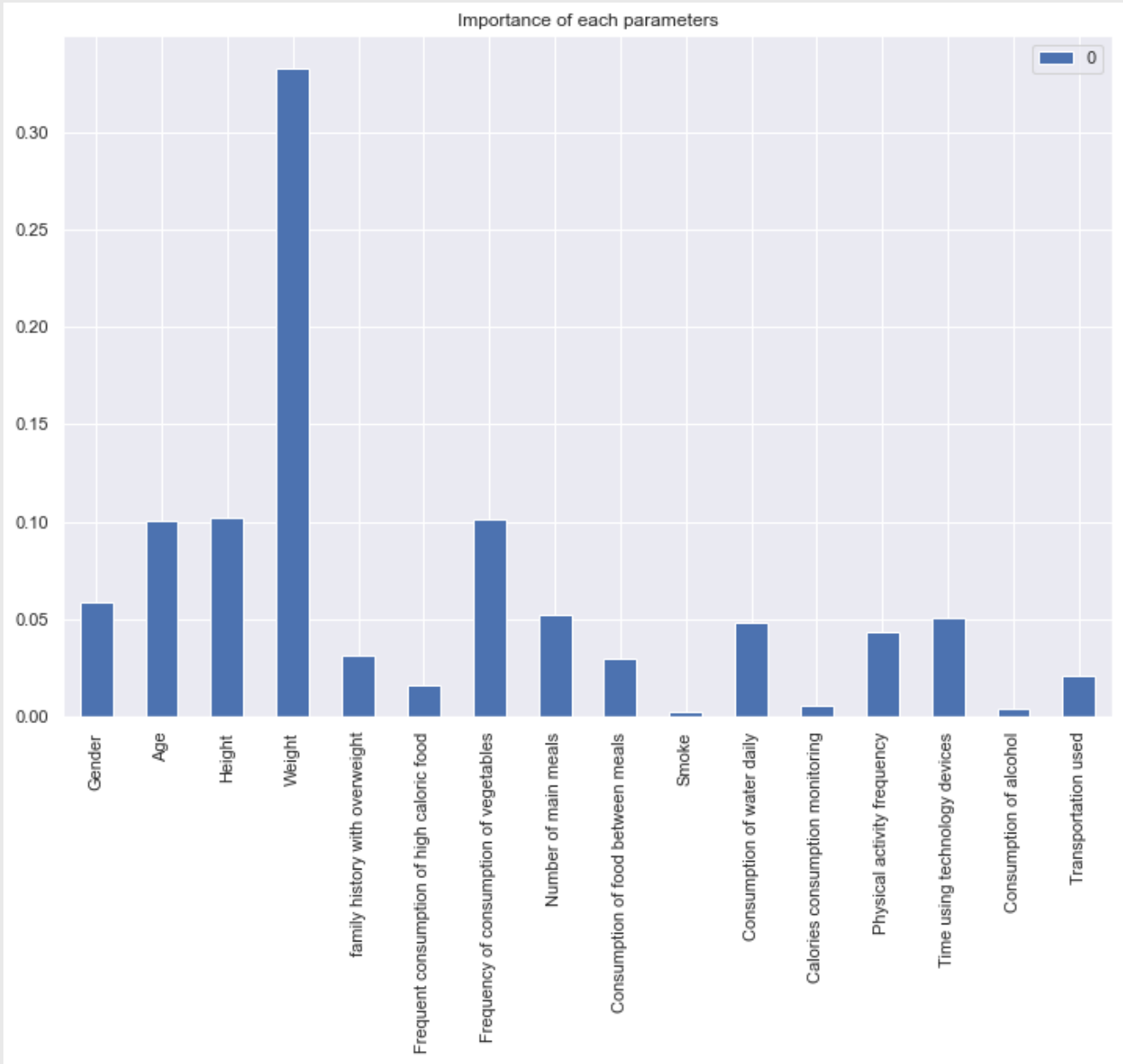


Parameters :

```
Gender                                   object
Age                                     float64
Height                                  float64
Weight                                  float64
family history with overweight           object
Frequent consumption of high caloric food object
Frequency of consumption of vegetables  float64
Number of main meals                    float64
Consumption of food between meals         object
Smoke                                     object
Consumption of water daily              float64
Calories consumption monitoring           object
Physical activity frequency             float64
Time using technology devices           float64
Consumption of alcohol                    object
Transportation used                       object
Obesity                                   object
dtype: object
```

Relation between obesity and parameters and Importance of parameters :





Problematic

Based on the habits and the physical condition, can we know if someone has a type of obesity and which type of it he has thanks to Machine learning algorithm ?

Machine learning :

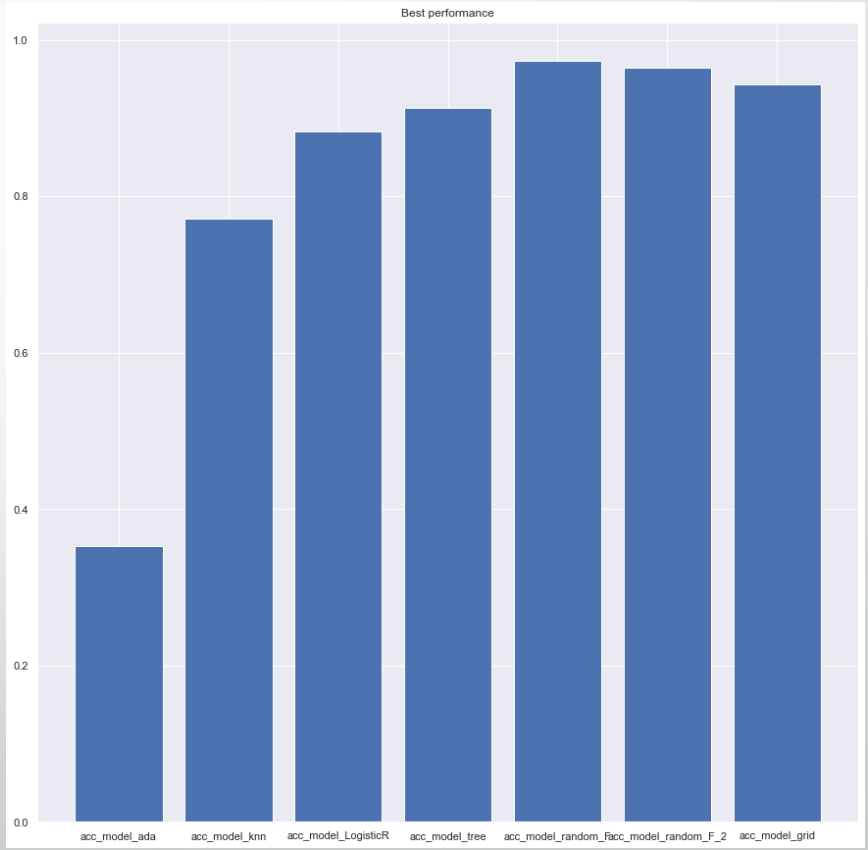
To predict Obesity with parameters, we use Machine learning.
So we need to prepare data for Machine leaning with these steps :

- Split of data set in : x_train, x_test, y_train, y_test
- Change qualitative variable to quantitative variable
- Scalling data

Models of Machine learning :

We use some models of Machine Learning :

- AdaBoost
- KNN (k-nearest neighbors algorithm)
- Logistic Regression
- Decision Tree Classifier
- Random Forest
- Grid Search
- With IMC instead of Weight and Height



=> Best performance with random forest

`Accuracy of : 0.9734848484848485`(with IMC)

API Flask :

We use flask to create an api of machine learning :

- With 4 of best predictors.

Example :