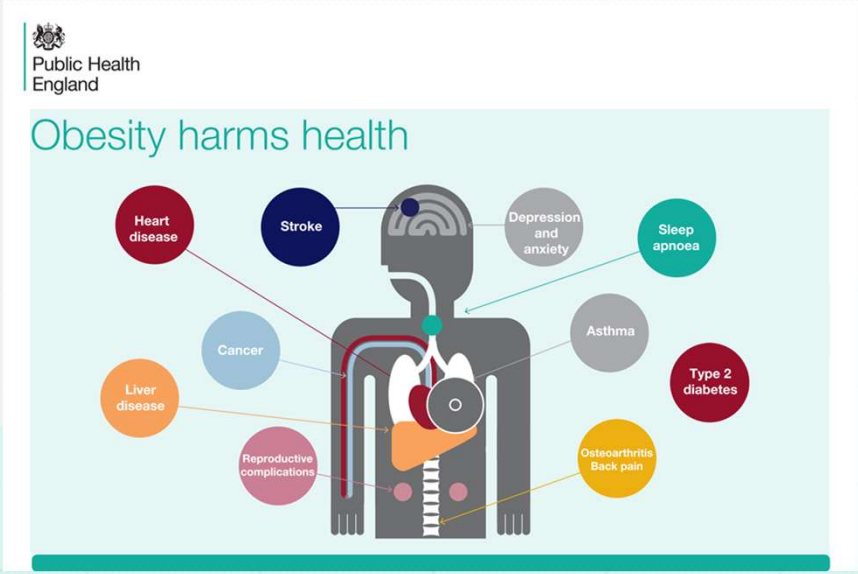
Estimation of obesity levels

Realised by Léopold Dumenil and Mohammed Hormi



Sommaire

1- Visualisation of the data

2- Analisis of the data

3- Prediction and Estimation with different models

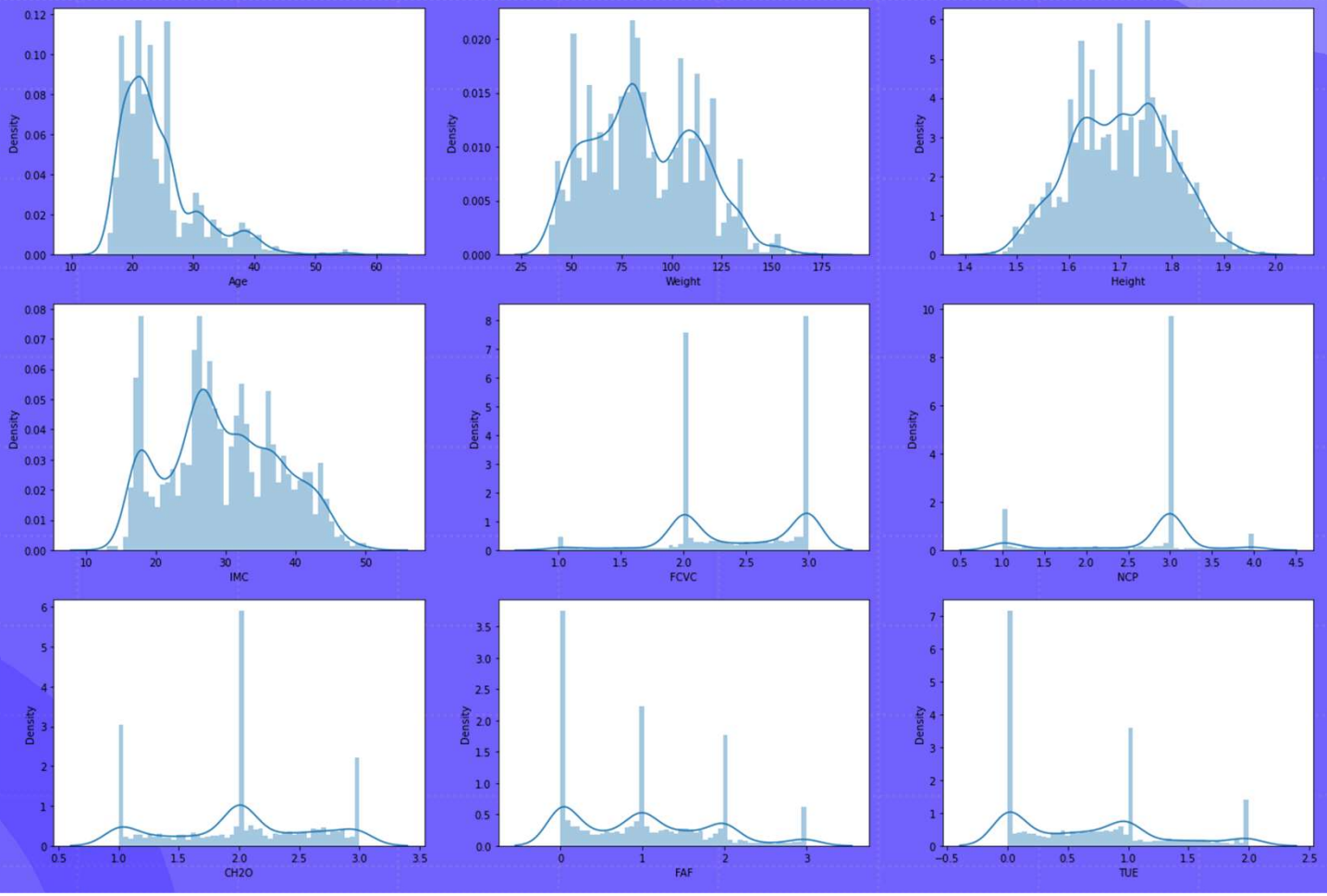
- a- Random forest
- b- Decision tree
- c- KNN
- d- Support Vector

4- Hyperparameters tuning

Visualisation of the data

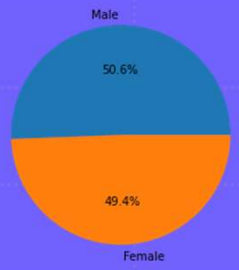
- First we saw with the function info() that there is no null data in our dataset so we don’t have to clean it.
- Moreover, with describe() we can see the statistics of the data and we see that the population is very young with a mean of age of 24,3 years old.
- After that, we decided to create a column BMI = Height / Weight^2

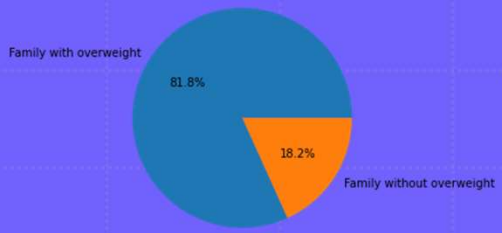
Visualisation of numerics data

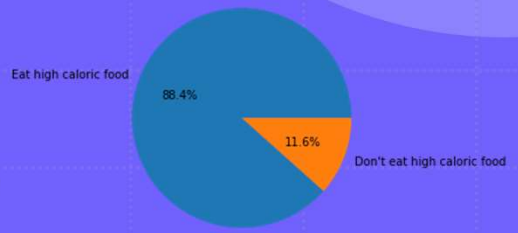


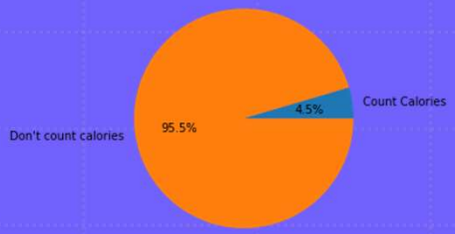
- Using these data representations, we can tell several things. First, the variables "Age", "Weight", "Height" and "BMI" are relatively well distributed over a wide range of values.
- However, we can see that the variable "NCP" which represents the number of meals per day is not very significant because the vast majority of people eat 3 meals. This variable will therefore be almost useless to create links with obesity cases.

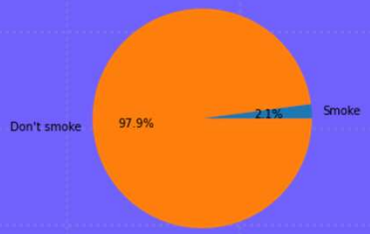
Visualisation of non-numerics data

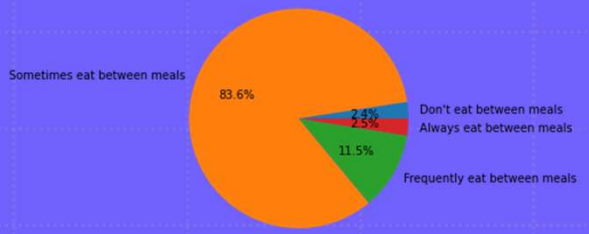


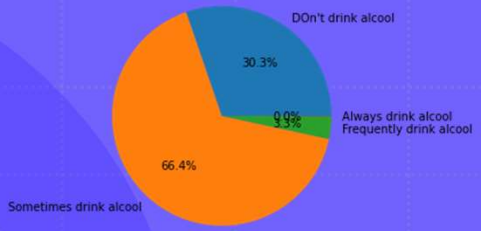


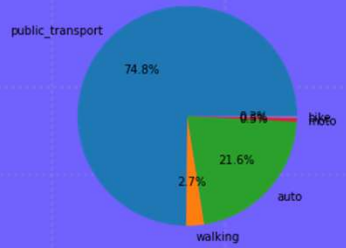


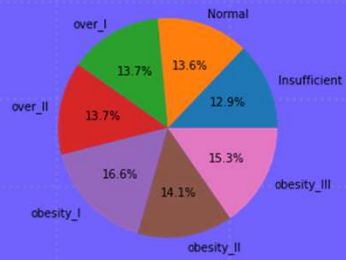






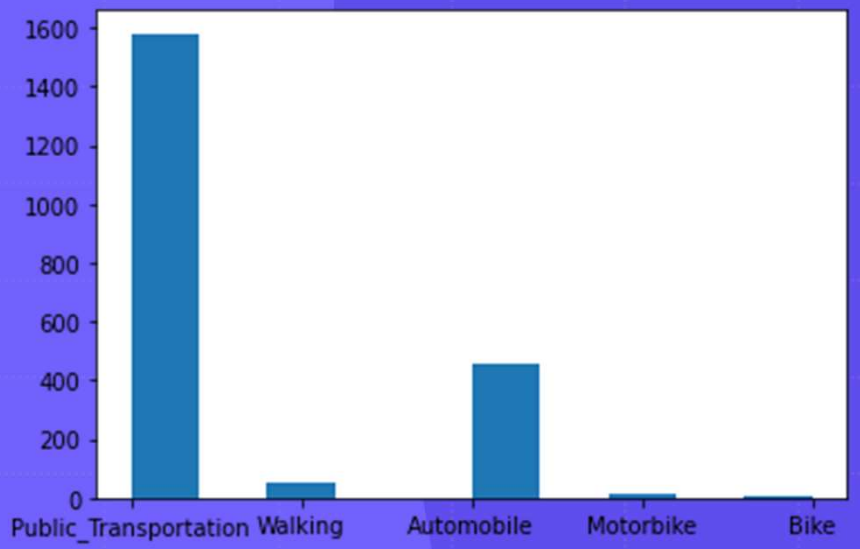






- As seen previously, some variables are not very useful because they are not well represented. Indeed, this is the case for smokers, there are too few smokers compared to non-smokers for this variable to be meaningful. The same is true for those who count or not their calories.
- What is interesting is that the "NObeyesdad" variable is relatively well distributed, each body type is represented with approximately the same proportionality. This distribution suggests that smoking or counting calories does not directly influence obesity.

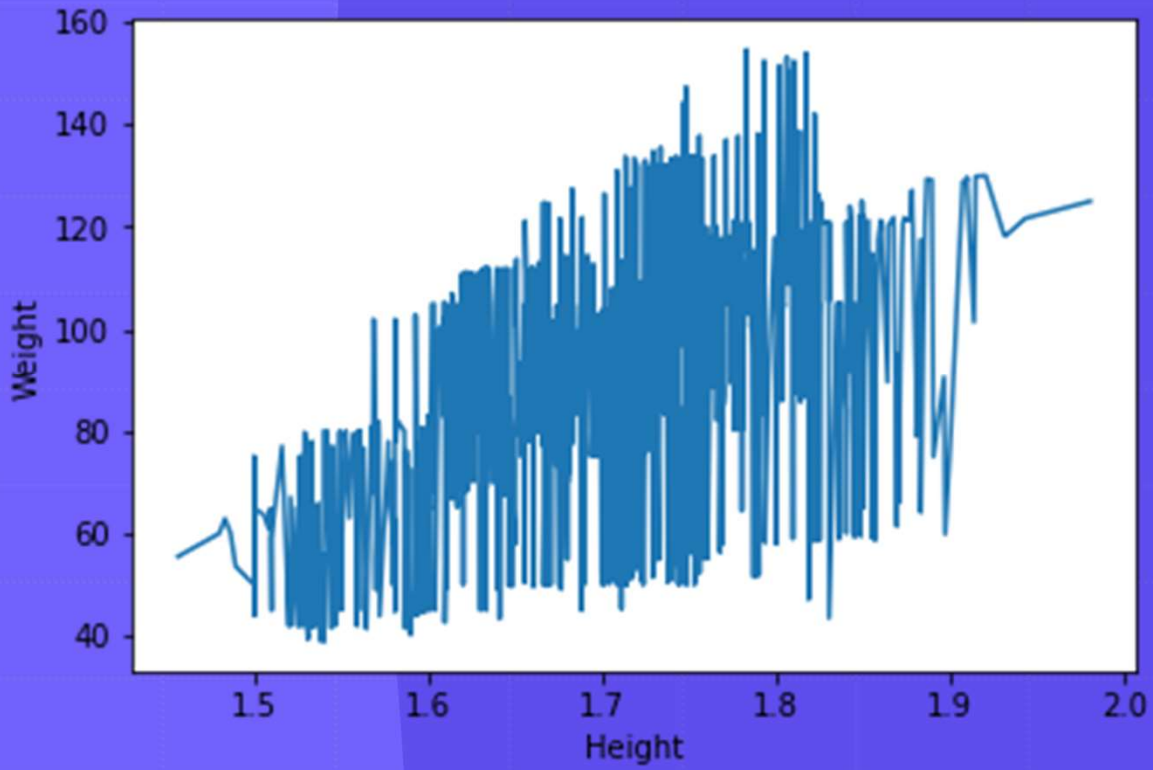
Other way to visualise data



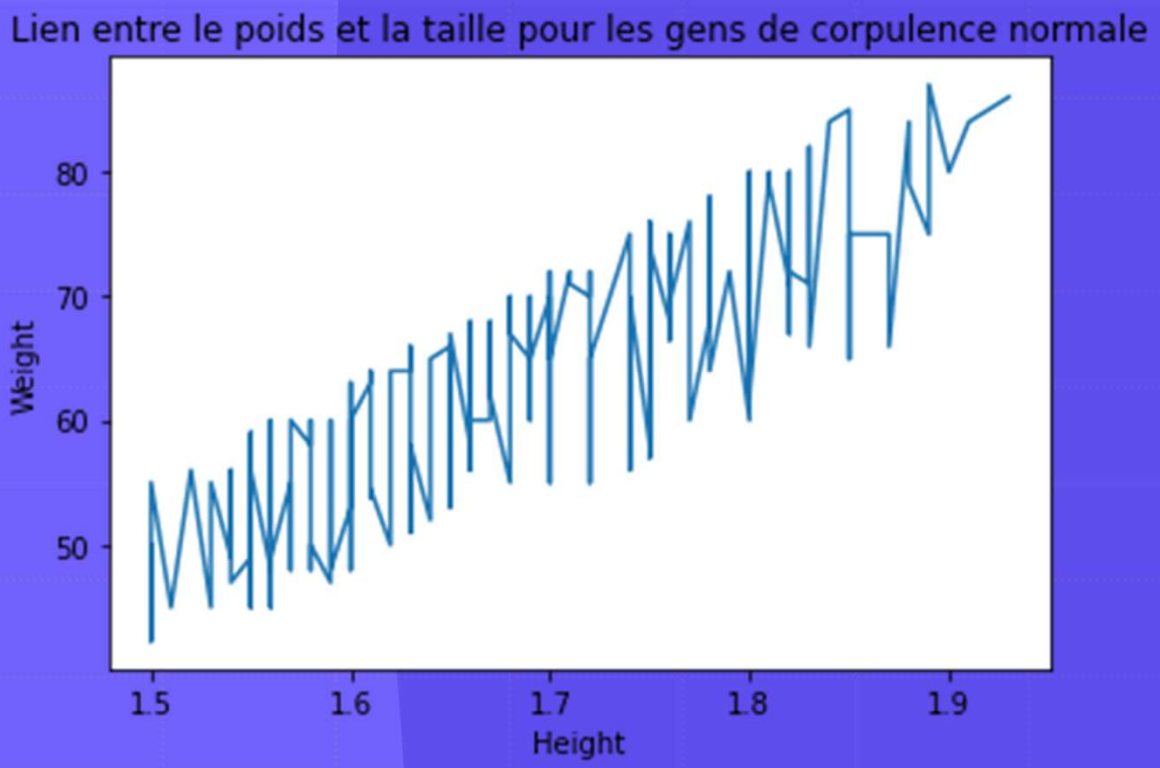
Plot with the number of the column MTRANS

Analysis of the data

- For this plot, I removed half of the data randomly so that it is a little more readable. We deduce that there is a link between height and weight (which is coherent). The tendency is increasing, i.e. the bigger you are, the heavier you are. However, the graph is difficult to use because many data do not follow the global trend. These data correspond to people whose weight is not "normal".

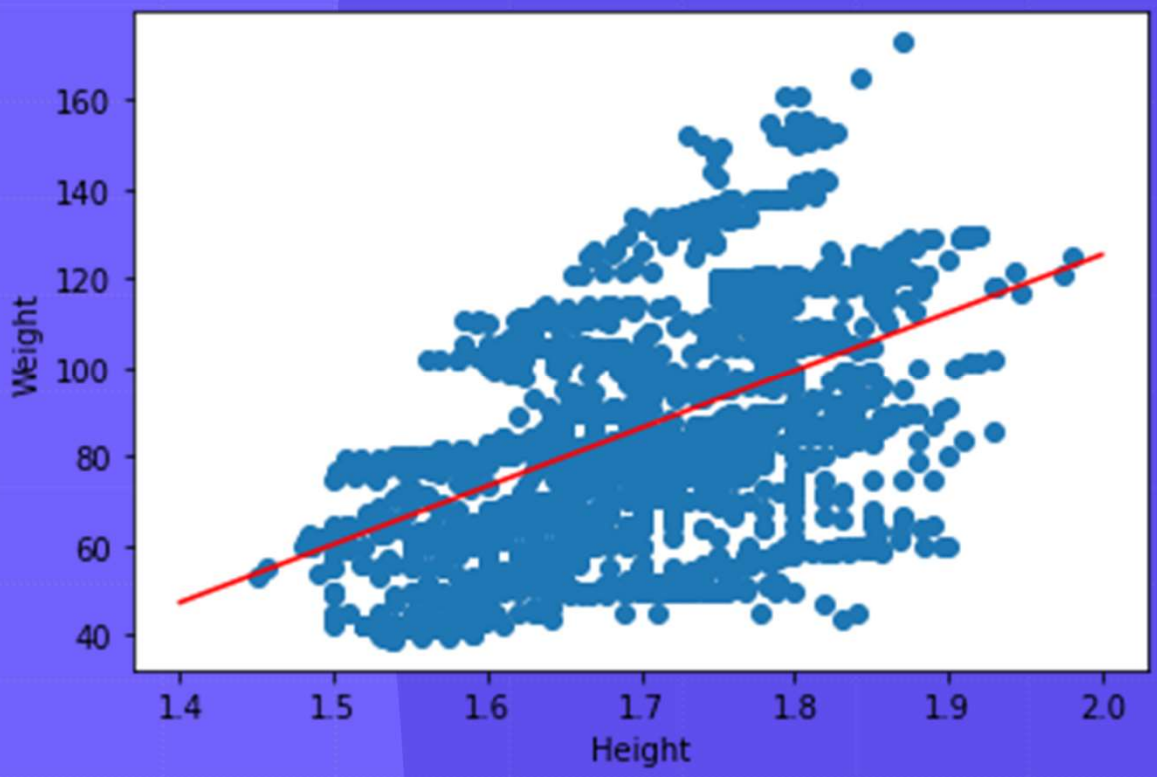


- Here is the plot if we take only the weight and the height of the people who have a value "Normal_Weight" for the variable "NObeyesdad". We can see that there are much less weird values and we can see a much more homogeneous curve.



Linear regression

- After linear regression with the module sklearn.linear_model, we can see the link between these two variables.



Relationship between differents columns

- BETWEEN BMI AND FCVC :
- The average BMI of people eating vegetables once a day is 23.61765536639728
- The average BMI of people who eat vegetables once every 2 days is 26.83102305656046
- The average BMI of people eating vegetables 1 time every 3 days is 31.00070278769006

- BETWEEN BMI AND FAF :
- Average BMI of people who exercise 0 times a week 30.120151201560727
- Average BMI of people who exercise 1 time per week 25.80191774727365
- Average BMI of people who exercise 2 times a week 22.516165760132022
- Average BMI of people who exercise 3 times a week 24.433015802050246

- BETWEEN BMI AND FAMILY_WITH_OVERWEIGHT_HISTORY :
- The average BMI of people with a family history is 31.529168764517884
- The average BMI of people with no history is 21.500493230325795

Conclusion of the analysis

- With these last three calculations, I wanted to check the influence of eating vegetables, doing sports and having a family history on BMI.
- In the first study, the results are obvious and show the influence of vegetables on body size. A person who eats vegetables regularly will have a BMI around 23.6 compared to 31 (obesity) for someone who rarely eats them.
- In the study on sports, the graph shows that the more one practices sports, the more one's BMI decreases and therefore the risk of obesity decreases. However, we notice an increase in BMI when we practice a lot of sport, this is probably due to the increase in muscle mass and not to fat mass.
- Finally, we notice that a family history of obesity favors obesity in people. Either it is simply a question of family culture with tendencies to eat a diet that is not fat or it could be a genetic criterion in which case people are born with a greater chance of becoming obese.

Data prediction with Random Forest

- We have an accuracy of 82.18, that is a good result.

```
Random Forest:
----------------------------------------------------------------
Accuracy: 0.82177
Accuracy w/Scaled Data (ss): 0.82177
Accuracy w/Scaled Data (mm): 0.82177

Classification Report (mm):
                     precision    recall  f1-score   support

Insufficient_Weight       0.85      0.87      0.86        92
      Normal_Weight       0.60      0.69      0.64        77
     Obesity_Type_I       0.85      0.80      0.82       114
    Obesity_Type_II       0.90      0.94      0.92        85
   Obesity_Type_III       0.99      0.99      0.99        92
 Overweight_Level_I       0.79      0.71      0.75        89
Overweight_Level_II       0.76      0.74      0.75        85

           accuracy                           0.82       634
          macro avg       0.82      0.82      0.82       634
       weighted avg       0.83      0.82      0.82       634

                          ---------------------------------
```

Data prediction with Desicion tree

- We have an accuracy of 76,5% that is less than the random forest method.

```
Decision Tree:
------------------------------------------------------------------
Accuracy: 0.76498
Accuracy w/Scaled Data (ss): 0.74763
Accuracy w/Scaled Data (mm): 0.75237

Classification Report:
                     precision    recall  f1-score   support

Insufficient_Weight       0.83      0.85      0.84        92
      Normal_Weight       0.56      0.48      0.52        77
     Obesity_Type_I       0.80      0.76      0.78       114
    Obesity_Type_II       0.88      0.84      0.86        85
   Obesity_Type_III       0.98      0.99      0.98        92
 Overweight_Level_I       0.74      0.65      0.69        89
Overweight_Level_II       0.56      0.74      0.64        85

           accuracy                           0.76       634
          macro avg       0.76      0.76      0.76       634
       weighted avg       0.77      0.76      0.77       634

Confusion matrix :
 [[78  7  1  0  0  0  6]
 [10 37  3  4  0 11 12]
 [ 4  4 87  2  1  6 10]
 [ 0  0  3 71  0  0 11]
 [ 0  0  0  0 91  0  1]
 [ 1 15  4  1  0 58 10]
 [ 1  3 11  3  1  3 63]]
                                  ----------------------------------
```

Data prediction with KNN

- We have an accuracy of 76,7% that is still less than random forest.

```
KNN:
------------------------------------------------------------
Accuracy: 0.76656
Accuracy w/Scaled Data (ss): 0.73502
Accuracy w/Scaled Data (mm): 0.73502

Classification Report:
                     precision    recall  f1-score   support

Insufficient_Weight       0.79      0.83      0.81        92
      Normal_Weight       0.71      0.35      0.47        77
     Obesity_Type_I       0.74      0.81      0.77       114
    Obesity_Type_II       0.79      0.98      0.87        85
   Obesity_Type_III       0.86      1.00      0.92        92
 Overweight_Level_I       0.76      0.67      0.71        89
Overweight_Level_II       0.67      0.66      0.66        85

           accuracy                           0.77       634
          macro avg       0.76      0.76      0.75       634
       weighted avg       0.76      0.77      0.75       634

Confusion matrix :
 [[76  2  4  1  0  2  7]
 [18 27  6  3  5  9  9]
 [ 1  3 92  6  1  5  6]
 [ 0  0  0 83  0  1  1]
 [ 0  0  0  0 92  0  0]
 [ 1  6 10  4  3 60  5]
 [ 0  0 13  8  6  2 56]]
                                  ----------------------------------
```

Data prediction with support vector machine

- Here we have a bad accuracy of 47,3%

```
SVM:
----------------------------------------------------------------
Accuracy: 0.47319
Accuracy w/Scaled Data (ss): 0.71609
Accuracy w/Scaled Data (mm): 0.71609

Classification Report (mm):
                     precision    recall  f1-score   support

Insufficient_Weight       0.79      0.80      0.80        92
      Normal_Weight       0.54      0.66      0.60        77
     Obesity_Type_I       0.65      0.58      0.61       114
    Obesity_Type_II       0.70      0.98      0.82        85
   Obesity_Type_III       0.98      0.98      0.98        92
 Overweight_Level_I       0.64      0.49      0.56        89
Overweight_Level_II       0.70      0.54      0.61        85

           accuracy                           0.72       634
          macro avg       0.71      0.72      0.71       634
       weighted avg       0.72      0.72      0.71       634

                    ----------------------------------
```

Now, we are going to change the hyperparameters using KNN

- Changing the hyperparameters, we have the best accuracy with 81%.
- Moreover, the values in the confusion matrix are better than in the normal KNN.

Classification Report:

	precision	recall	f1-score	support
Insufficient_Weight	0.83	0.84	0.83	92
Normal_Weight	0.76	0.51	0.61	77
Obesity_Type_I	0.83	0.83	0.83	114
Obesity_Type_II	0.85	0.94	0.89	85
Obesity_Type_III	0.91	1.00	0.95	92
Overweight_Level_I	0.78	0.78	0.78	89
Overweight_Level_II	0.70	0.75	0.73	85
accuracy			0.81	634
macro avg	0.81	0.81	0.80	634
weighted avg	0.81	0.81	0.81	634

Confusion matrix :

```
[[77  4  3  0  0  3  5]
 [12 39  4  1  4  7 10]
 [ 1  3 95  3  1  5  6]
 [ 0  0  1 80  0  1  3]
 [ 0  0  0  0 92  0  0]
 [ 3  5  6  3  0 69  3]
 [ 0  0  6  7  4  4 64]]
```

Conclusion

- To conclude, the best method is KNN changing the hyperparameters with an accuracy of 81%.

- For the prediction, we have decided to drop the columns Height and Weight because they are directly involve in the obesity level with BMI so it was not interesting to keep them in the data.