Introduction

This presentation is about a dataset for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition.

To start we will introduce the dataset and its features. Then we will present how did we perform this study.

The following is a general overview of our study, the thinking process and the methods employed. For the full study, you should check our notebook.

The dataset

The data :

- It contains 2111 records.
- 77% of the data was generated synthetically.

The features :

- There are 17 features.
- The response variable is “NObesity” which represents the obesity level of an individual. These variable has seven categories from insufficient to obesity level 3.

The dataset

Which features are categorical?
These values classify the samples into sets of similar samples. Within categorical features are the values nominal, ordinal, ratio, or interval based? Among other things this helps us select the appropriate plots for visualization.

Categorical:
- Gender
- family_history_with_overweight
- FAVC wich is the consumption of high caloric food
- SMOKE
- SCC
- MTRANS wich is the mode of transport

Interval-based:
- CH20 which is how much water drink per day
- FAF which is the the physical activity frequency
- TUE which the time using technology devices

Ordinal:
FCVC which represents the frequency of consumption of vegetables
CAEC which is the consumption of food between meals
CACL which is how often the person drink alcohol
NObeyesdad which is the obesity level

The dataset

Which features are numerical?
These values change from sample to sample. Within numerical features are the values discrete, continuous, or timeseries based? Among other things this helps us select the appropriate plots for visualization.

Continuous:
- Age
- Height
- Weight

Discrete :
- NCP which is the number of meal per day

Statistical Analysis

We started to work on this data by performing a simple statistical analysis. We did that in order to have a global comprehension of the data.
The following results are quite interesting and allowed us to make our first assumptions about which features will be used in the models.

```
Proportion of people who smoke : SMOKE
no       97.968856
yes       2.031144
```

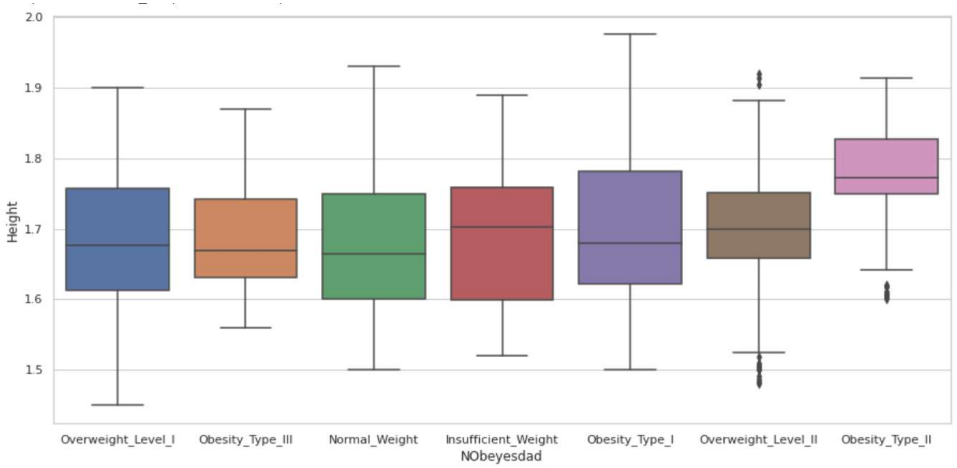
- Almost **98%** of the dataset do not smoke. This feature do not seem to provide information because this variance is close to zero.

```
Proportion of people who monitor calories consumptions
no     95.734597
yes     4.265403
```

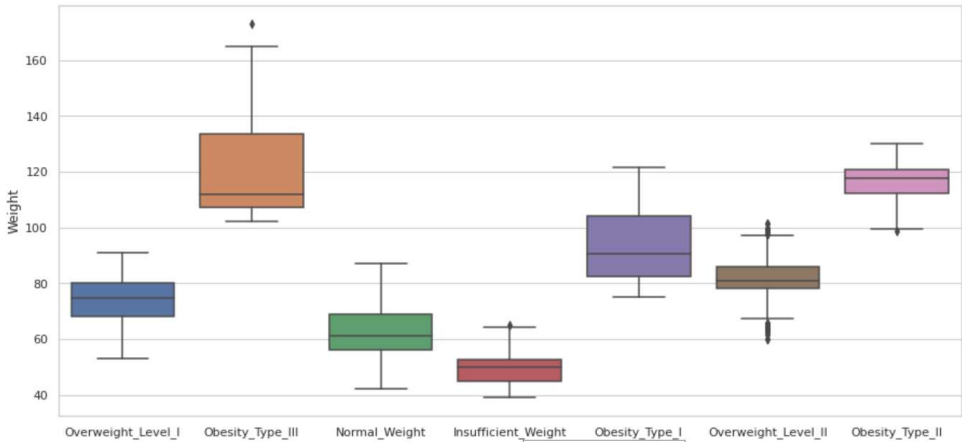
- Same case with this other feature. Almost **96%** of the dataset do not monitoring their calories consumption.

Features “Height” and “Weight”

People commonly think that the obesity level of a person is linked to his height and to his weight.
In order to assess or not this statement, we started our detailed analysis by studying the features “Height” and “Weight”.



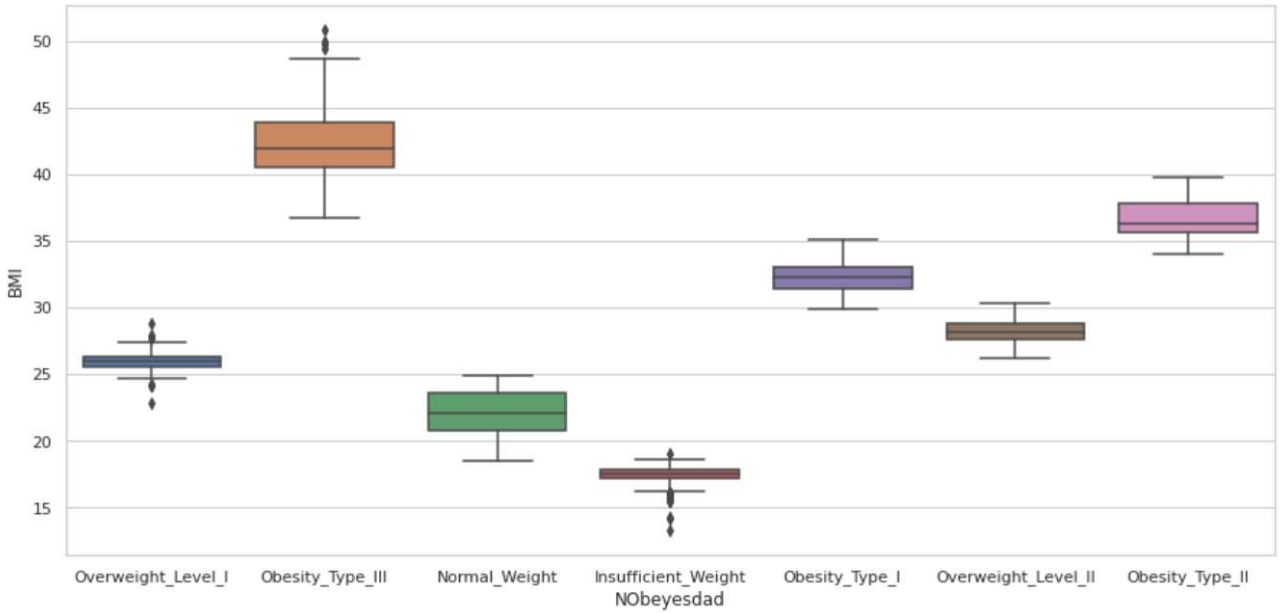
The feature **“Height”** does not seem as relevant as we used to think.



The feature **“Weight”** seems very relevant. Each obesity level has its own range of weight, well separated from the other.

Features “Height” and “Weight”

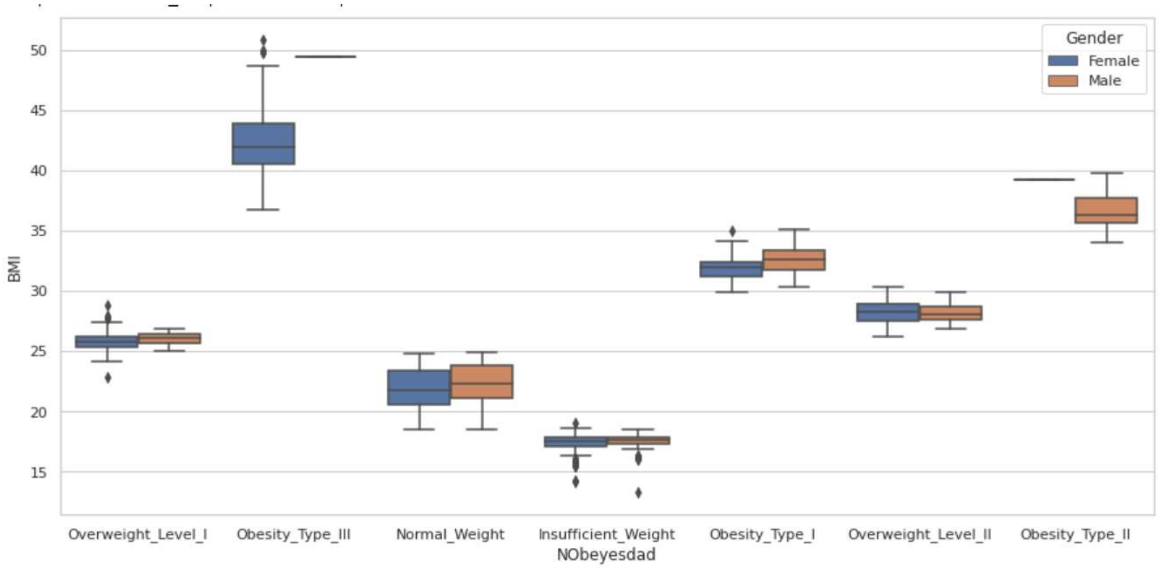
We observed the outliers of the feature “Height” and we noticed that the people which had a height greater than the others also had a weight greater than the others.
We decided to study the relationship between these two features then we modelized it by creating the **BMI** feature which is the **body mass index.** The following is the formula of the BMI : $(Weight * Height)/(Height^2)$



- The relationship between this new feature and the response variable is very strong.
- The relationship between this new feature and the response variable is stronger than the relation between the variable “Weight”/”Height” and the variable response.

Feature “Gender”

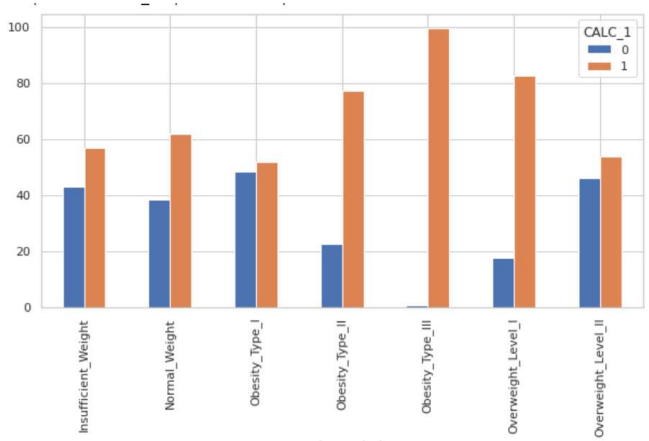
We observed the outliers of the feature “BMI” and we noticed that all outliers were woman. We decided to check if this feature is linked to the feature “BMI” and to the variable response.



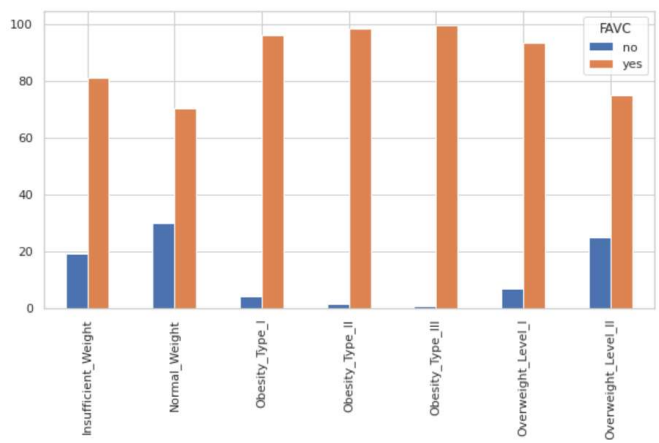
- Visually, the link between these features was not clearly establish. However, we noticed that for each obesity level, women have a mean weigh inferior to men.
- We deeply think that the gender is important in the prediction of the obesity level. The interpretation of the BMI depends on the gender.
- We run a chi-2 test to confirm our assumption and according to the test, we retained this feature.

Features about the eating habits

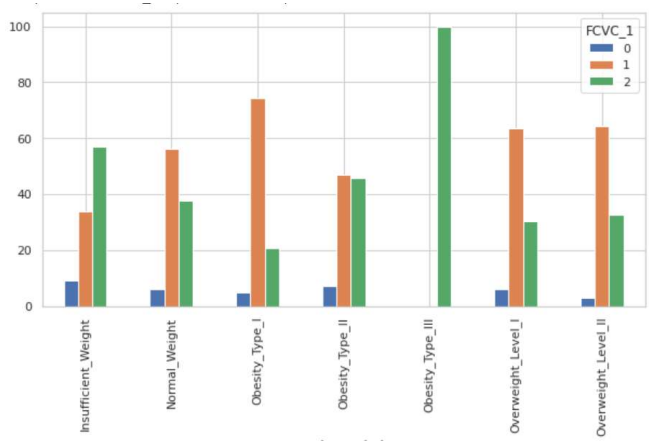
We decided to gather the analysis on the features about the eating habits in order to combine some of them. First, we modified the feature individually and we obtained the following features:



This is **the consumption of alcohol** according to the level of obesity. We merge some categories into a more general one. We did that some categories were underrepresented.



This is **the consumption of high caloric food** according to the level of obesity.

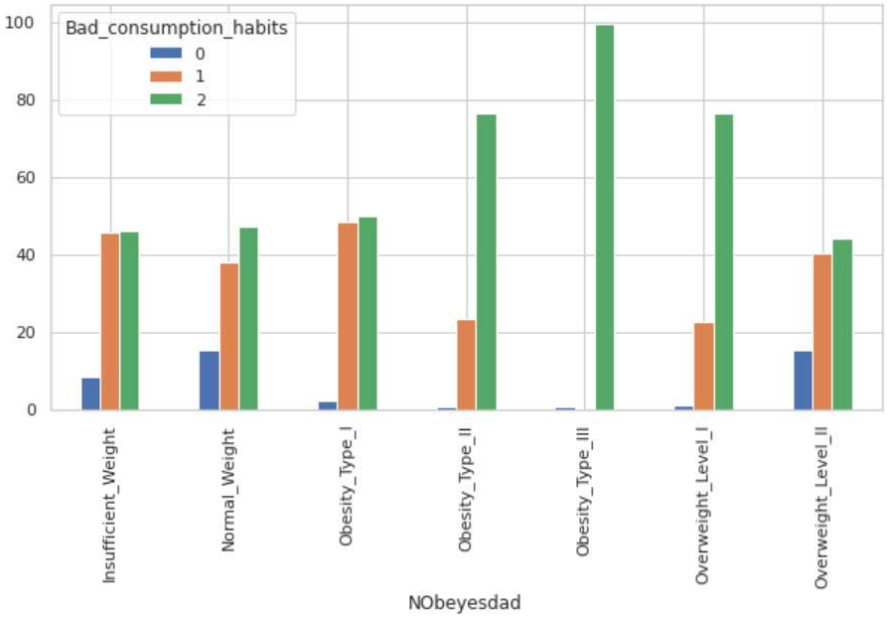


This is **the consumption of vegetables** according to the level of obesity.

Features about the eating habits

These three features contains relevant information. For example, we established from the feature “CALC” that people with the most severe kind of obesity drinks much alcohol than the others.

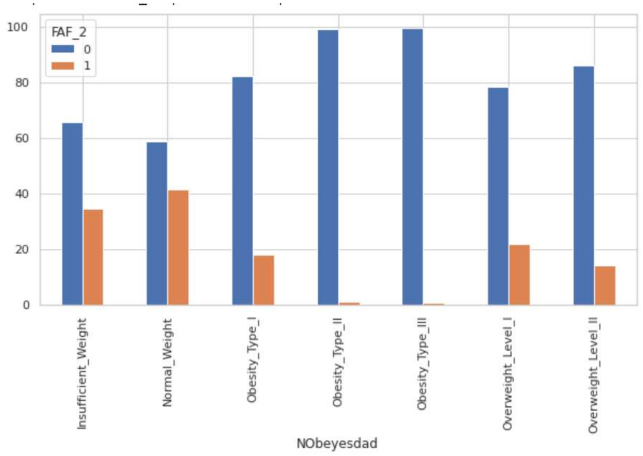
We decided to combine these features because they are complementary. We decided to create the feature **“Bad consumption habits”.** It is a scale which follows the rules : +1 if the person drinks alcohol, +1 if the person eats high caloric food and +1 if the person does not eat vegetables.



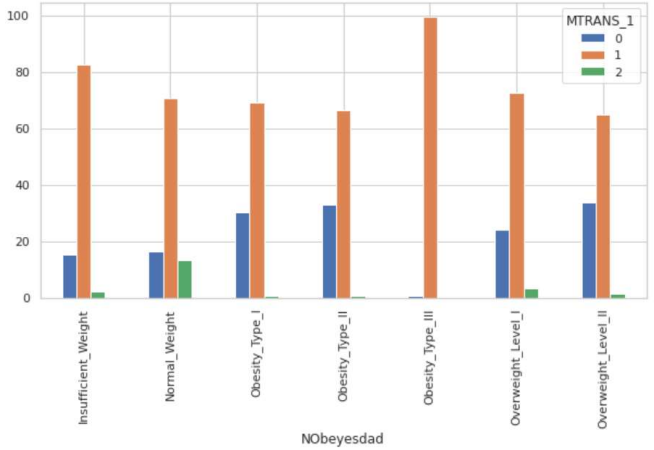
- This feature seems to be a good sum up.
- We compared it with the previous ones thanks to a chi-2 test. This feature is much relevant than the feature about the alcohol consumption and the one about the high caloric food consumption but less than the feature about the vegetables.
- We drop the features about the vegetables from the feature “Bad consumption habits”. And we used them both for our models.

Features about the sport habits

We decided to gather the analysis on the features about the sport habits in order to combine some of them. First, we modified the feature individually and we obtained the following features:



This is the feature about the frequency of physical activity (FAF). In order to make this feature more relevant, we set at 1 if a person respects the OMS' recommendations about physical activity

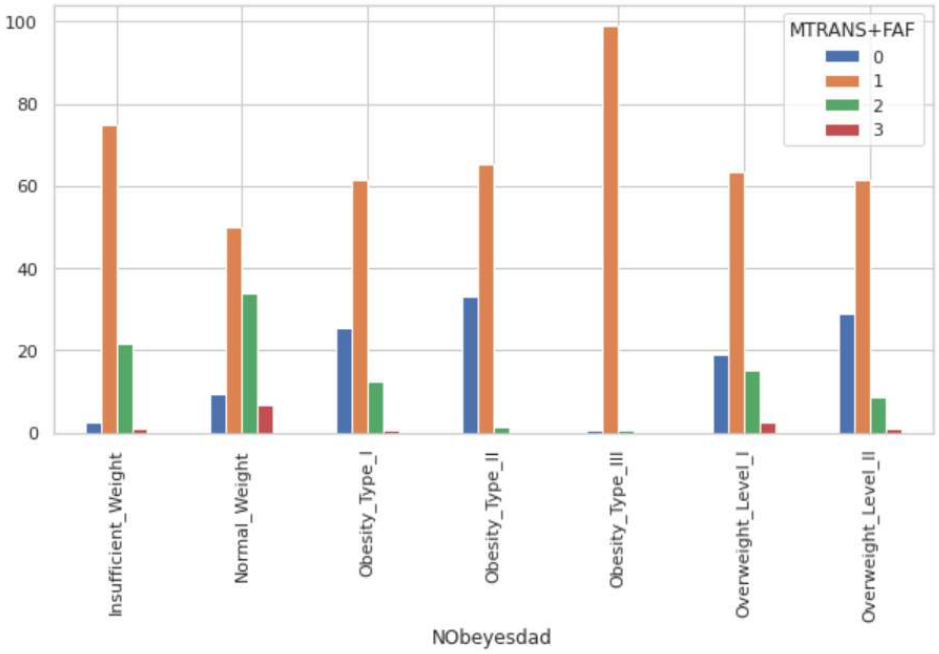


This is the feature about the means of transportation. In order to combine this feature with the feature FAF we made some changes.

- "Walking" and "Bike" which imply a physical activity are gathered in the same category.
- "Motorbike" and "Automobile" which do not imply physical activity are gathered too.

Features about the sport habits

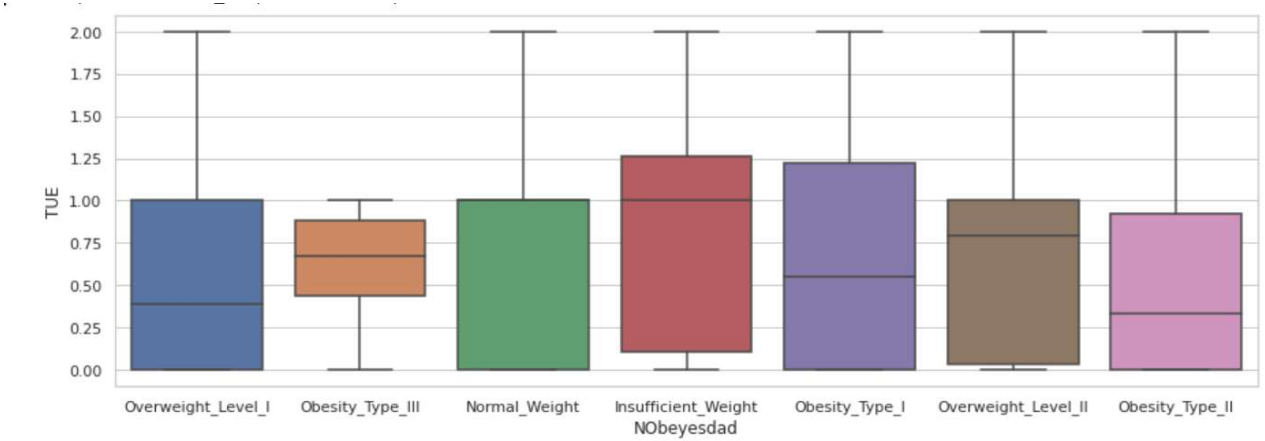
We believe that the feature about the means of transportation contains some information but not enough to be relevant by itself. We decided to combine this feature with the feature “FAF”. Both are about physical activity so it could be relevant. The new feature is the addition of them.



- This features is correlated with the variable response.
- It clearly appeared that people who do not have a normal weight perform less physical activity than the others.
- We confirmed that this feature is more relevant than the features is made from by computing a chi-2 test.

Feature about the time using technology devices

The 'TUE' feature gives information on the daily usage time of technological devices. We asked ourselves if such a feature could be relevant in our study.
By first plotting a boxplot, we can predict that this feature will probably not be retained. Indeed, the use of devices does not seem to affect physical conditions.



To confirm this idea, we performed a Chi2 test on this feature. This test gives us a p-value equal to 0.07, therefore greater than 0.05. We therefore cannot reject the null hypothesis, ie the independence of the variables. We therefore chose to drop the TUE feature.

Models

We tried several models. We tried to experiment different kind of models:

- Discriminant Analysis models
- K-Nearest Neighbors model
- Tree based models
- Neural Network Classifier

We chose to use two metrics in order to evaluate these models :

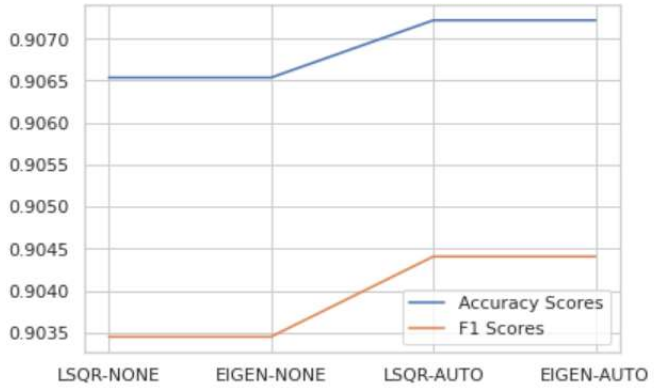
- the **accuracy**
- the **F1 score**.

These two metrics are commonly used in classification problem therefore we chose them.

Before fitting the models, we preprocessed the data by encoding the categorical variables.

Linear Discriminant Analysis

- Overview : We did not scale our data because there is no need to for this model.
- We implemented a grid search in order to test the different solvers available for this model. We also test different shrinkage methods.

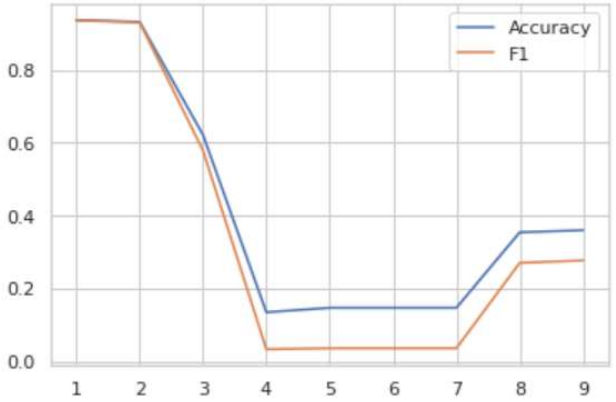


The best estimator occurred when we used the “LSQR” solver with an auto shrinkage or when we used the “EIGEN” solver with an auto shrinkage. The accuracy on the validation set are about **91% for both models**.

- When we obtained the best estimator, we decided to use a cross validated features selection method in order to keep only the most relevant features.
- We were finally able to evaluate the model. **Both metrics were equals to 90%** on the test set.

Quadratic Discriminant Analysis

- Overview : We did not scale our data because there is no need to for this model.
- We early encountered problem because some features were collinear. We decided to use a features selection method based on the chi-2 test. This method select only the K features which are the most correlated to the response variable. We applied this method for K from 1 to 9 and foreach K we fitted a quadratic discriminant analysis using cross validation.

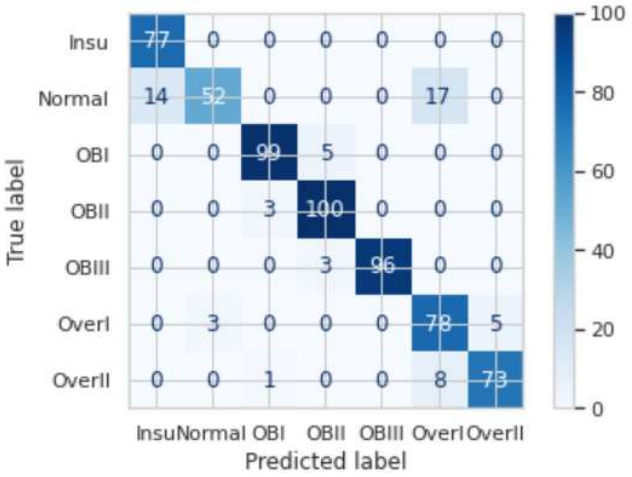


We obtained the best scores when the number of features is one or two. The scores on the validation set are slightly better when we only used one variable and this variable is the BMI.

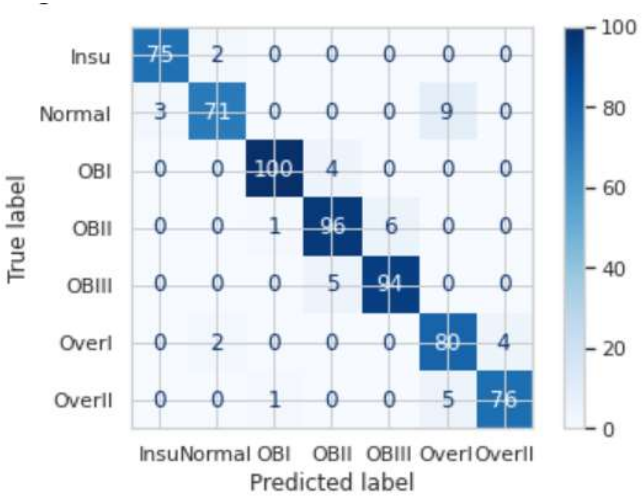
- We fitted the quadratic analysis on the train set with only the BMI feature. We did not use a grid search because we did not want to change the parameter.
- We were finally able to evaluate the model. **Both metrics were equals to 93%** on the test set.

Quadratic versus Linear

In order to compare the models with precision we decided to compute the confusion matrix.



Confusion Matrix of the linear model.

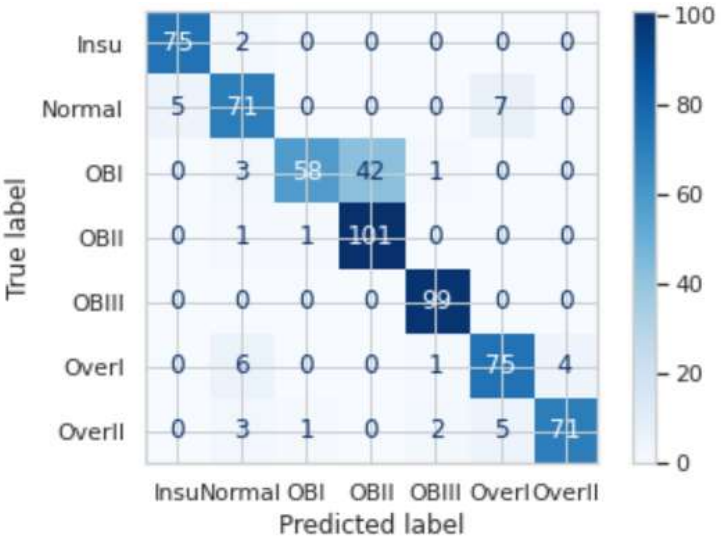


Confusion Matrix of the quadratic model.

- The linear model is unable to correctly distinguish the category “Normal Weight“ from the category “Insufficient Weight”. Same observation for the categories “Normal Weight” and “Overweight I”.
- The quadratic model fix the previous issues.
- The linear model is more able to correctly distinguish the category "Obesity II" from "Obesity III".

Naïve Bayes

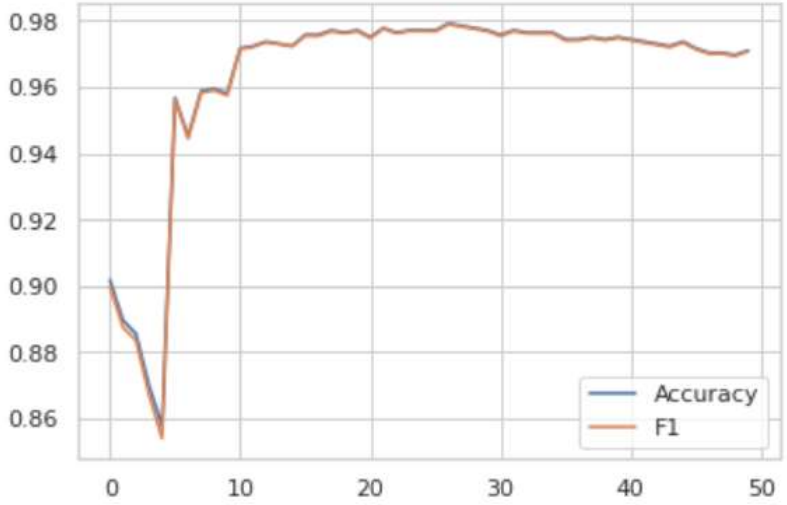
We tried to use Naive Bayes classifier to improve our results. However, this model did not result in sufficient precision even when using cross validation. Here is the confusion matrix we obtained:



We can see a significant number of errors on our test set. The accuracy is only 86.9%.

Bagging Classifier

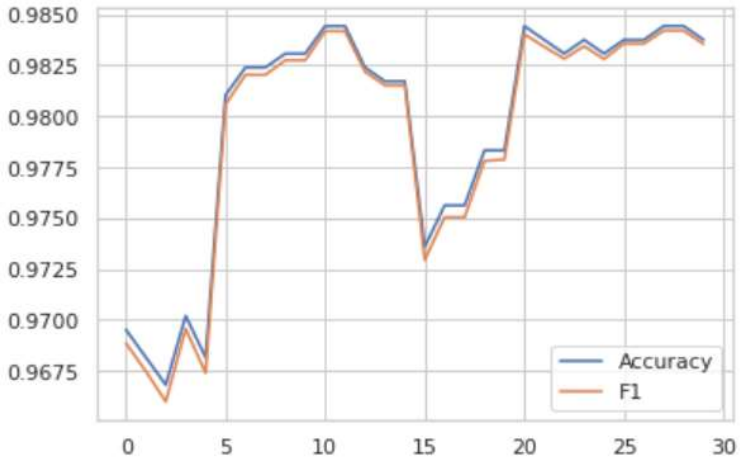
- Overview : We did not scale our data because there is no need to for this model.
- We implemented a grid search in order to find the best number of trees, the best criterion to make the split and the best number of features to select at each split.



The best estimator occurred when we used 200 trees, the entropy criterion and when we selected 6 features at each split. This model has an accuracy of 97.8% on the validation set.

Random Forest

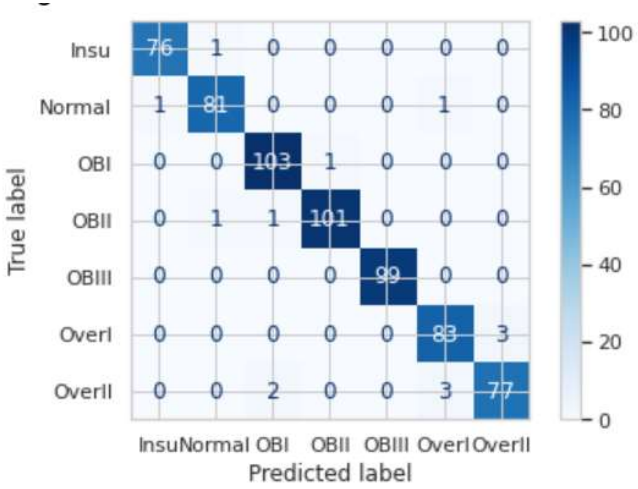
- Overview : We did not scale our data because there is no need to for this model.
- We implemented a grid search in order to find the best number of trees, the best criterion to make the split and the best number of features to select at each split.



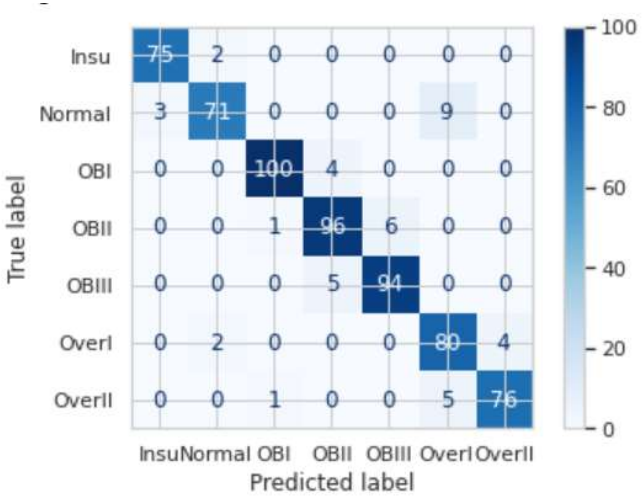
The best estimator occurred when we used 100 trees, the entropy criterion and when we selected 4 features at each split. This model has an accuracy of 98.4% on the validation set.

QDA versus Random Forest

In order to compare the models with precision we decided to compute the confusion matrix.



Confusion Matrix of the Random Forest.

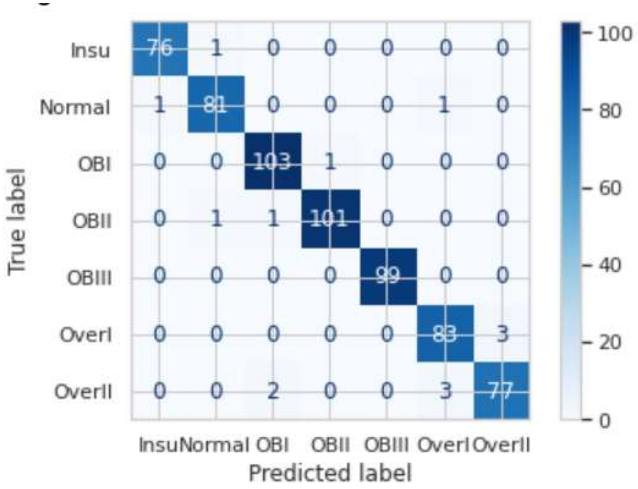


Confusion Matrix of the quadratic model.

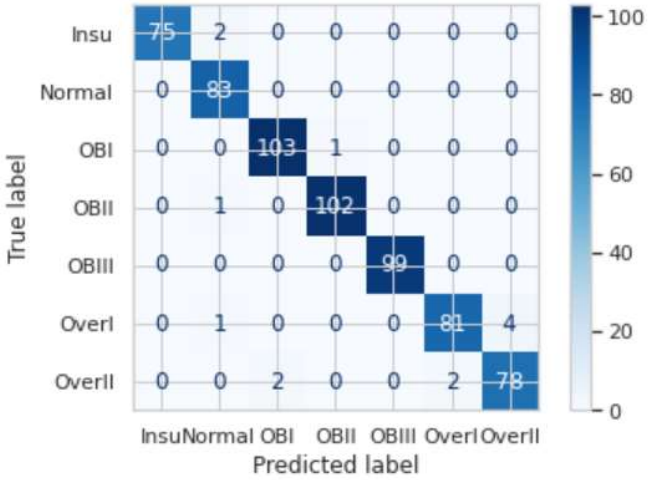
- Most of the problems encountered by the discriminant analysis models are fixed.
- The Random Forest is not perfectly able to distinguish the category “Overweight I” from the category “Overweight II”.
- This is the best model we obtained with a final accuracy score of 97.7 on the test set.

Random Forest versus Bagging Classifier

We decided to compare the Random Forest and the Bagging Classifier.



Confusion Matrix of the Random Forest.

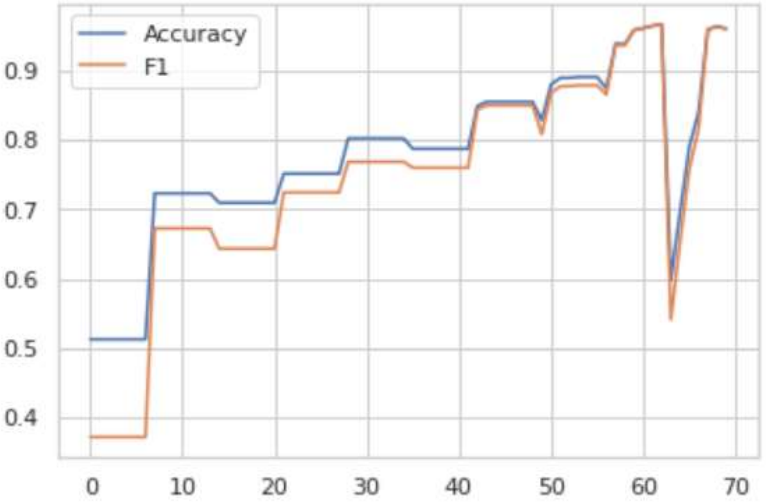


Confusion Matrix of the Bagging.

- We can observe that the Bagging Classifier is more precise than the Random Forest.
- This is also confirmed via their prediction: 97. 7 for the Random Forest against 97.8 for the Bagging

ADA Boost Classifier

- Overview : We did not scale our data because there is no need to for this model.
- We implemented a grid search in order to find the best number of estimators and the best learning rate. The default learning rate is 1 so, we chose ten values between 1 and 2. We tried different range of values for the number of estimators.

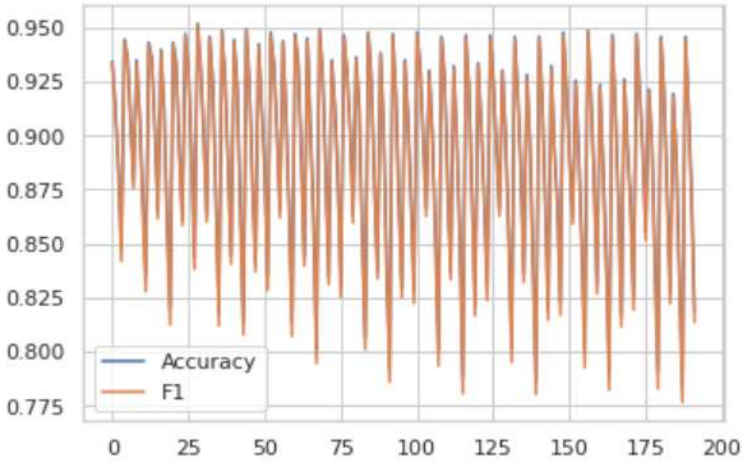


The best estimator occurred when we used 600 weak classifiers and when we set the learning rate at 1.89. This model has an accuracy of 96.4% on the validation set and an accuracy of 97% on the test set.

The value of these scores are lower than those we obtained with the Random Forest, so we did not make further analysis.

KNN Classifier

This method works poorly in high dimensions, this is because of the curse of dimensionality. Indeed, in high dimensions, measure of distance such as the Euclidean distance are meaningless. We need to reduce the dimensions before fitting the model.
This method is also scale sensitive so, we need to scale our data.

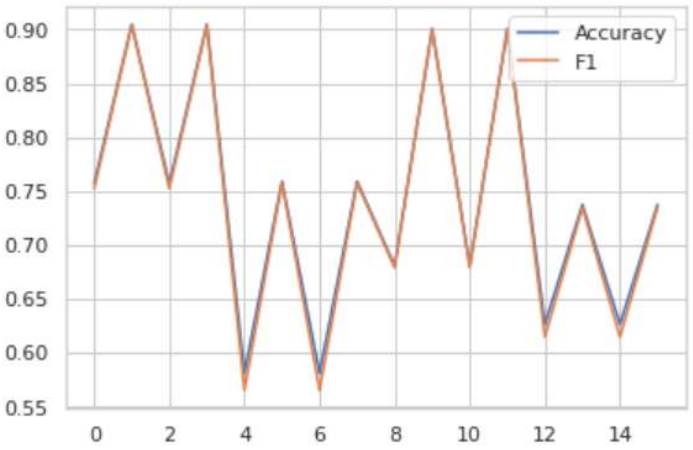


The best estimator is obtained when the number of components of the PCA is 2. We are in two dimensions so the Euclidean distance or the Manhattan distance can be used without doubt. This model has an accuracy of 95.2% on the validation set and 94.8% on the test set.

The value of these scores are lower than those we obtained with the Random Forest & Bagging Classifier, so we did not make further analysis.

MLP Classifier

MLPClassifier stands for Multilayer Perceptron Classifier which, in the name itself, connects to a neural network. Unlike other classification algorithms such as Naive Bayes Classifier, MLPClassifier relies on an underlying neural network to perform the classification task.



This model has an accuracy of 90.4% on the validation set and 88.9% on the test set.

The value of these scores are lower than those we obtained with the Random Forest & Bagging Classifier, so we did not make further analysis.