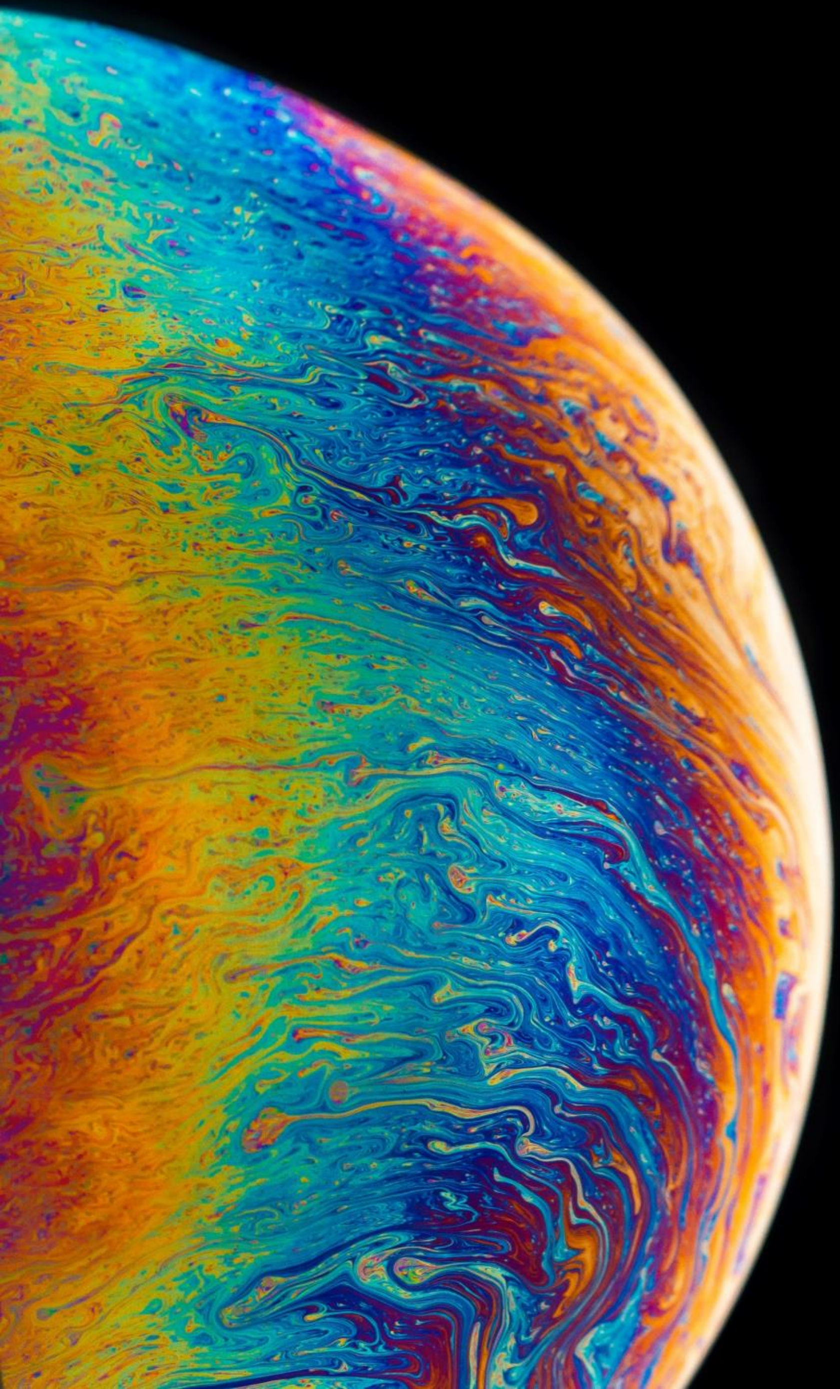


# Python for Data Analysis

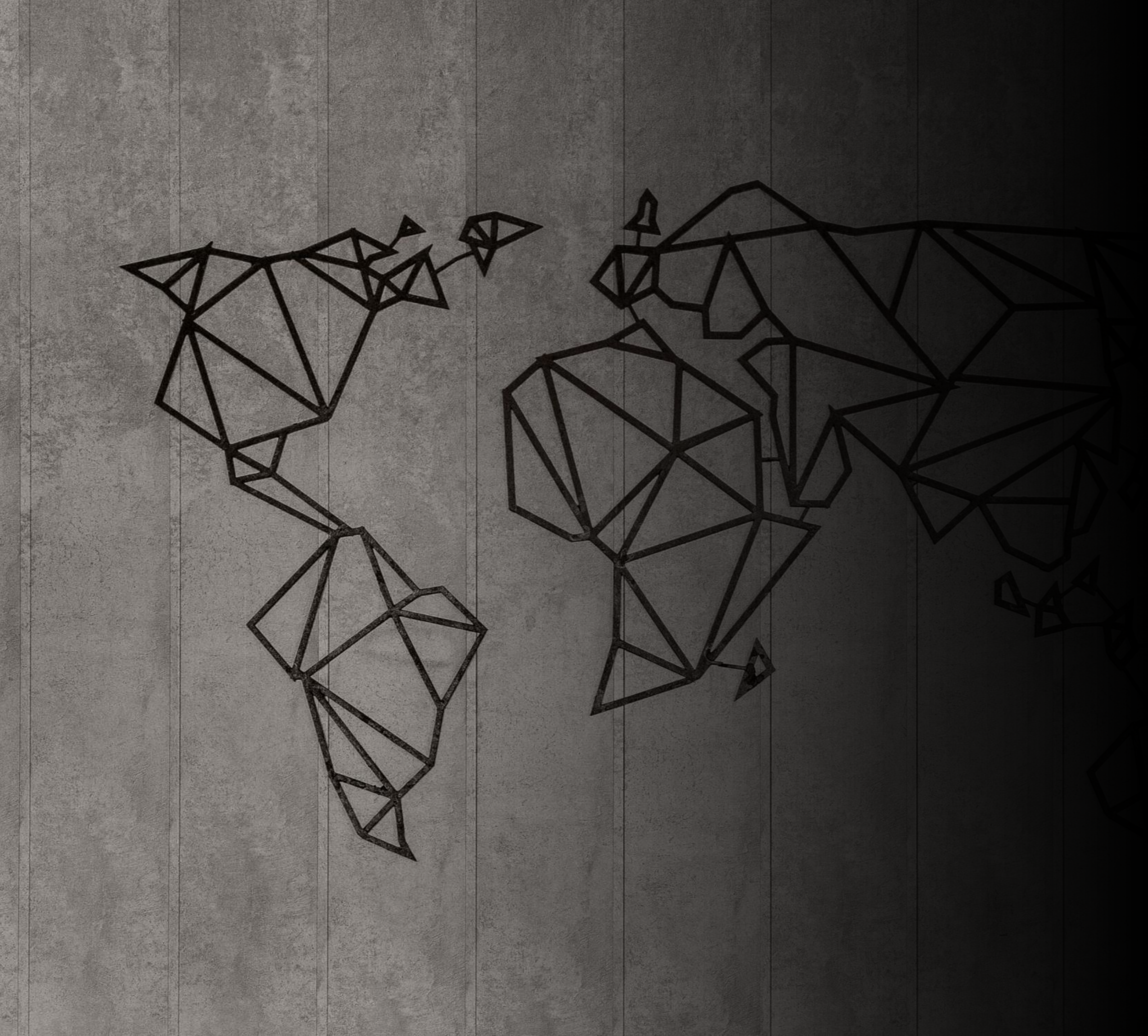
-

# Final project

*Louise Farré & Pierre-Victor Landez*

*-*

*DIA 4*



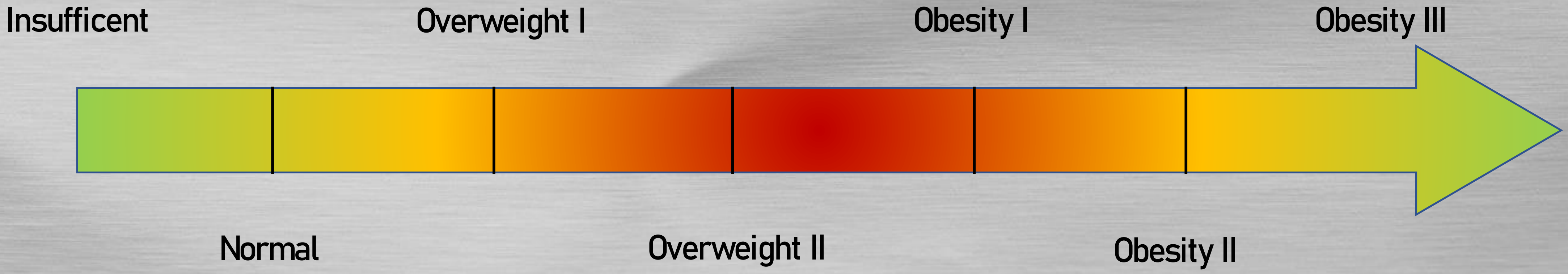
# Presentation of the project

- This study is about an estimation of obesity levels from population in Mexico, Peru and Columbia

- It is based on their eating habits and physical condition.

# Variable of interest : NObeyesdad

- There are 7 categories of weight from Insufficient to Obesity lvl 3



# The Survey

- The dataset is the result of a 16 questions-survey on a web platform
- 77% of the dataset was generated with Weka tool and the SMOTE filter

# The Questions

- ✓ *Age ?*
- ✓ *Gender ?*
- ✓ *Height ?*
- ✓ *Weight ?*
- ✓ *Any family history of overweight ?*
- ✓ *High caloric food ?*
- ✓ *Vegetables in **your** meals ?*
- ✓ *How many meals a day ?*
- ✓ *Eat food between meals ?*
- ✓ *Smoke ?*
- ✓ *Water consumption ?*
- ✓ *Do you count calories ?*
- ✓ *Physical activity ?*
- ✓ *Screen time per day ?*
- ✓ *Alcohol consumption ?*
- ✓ *Most frequent transportation ?*

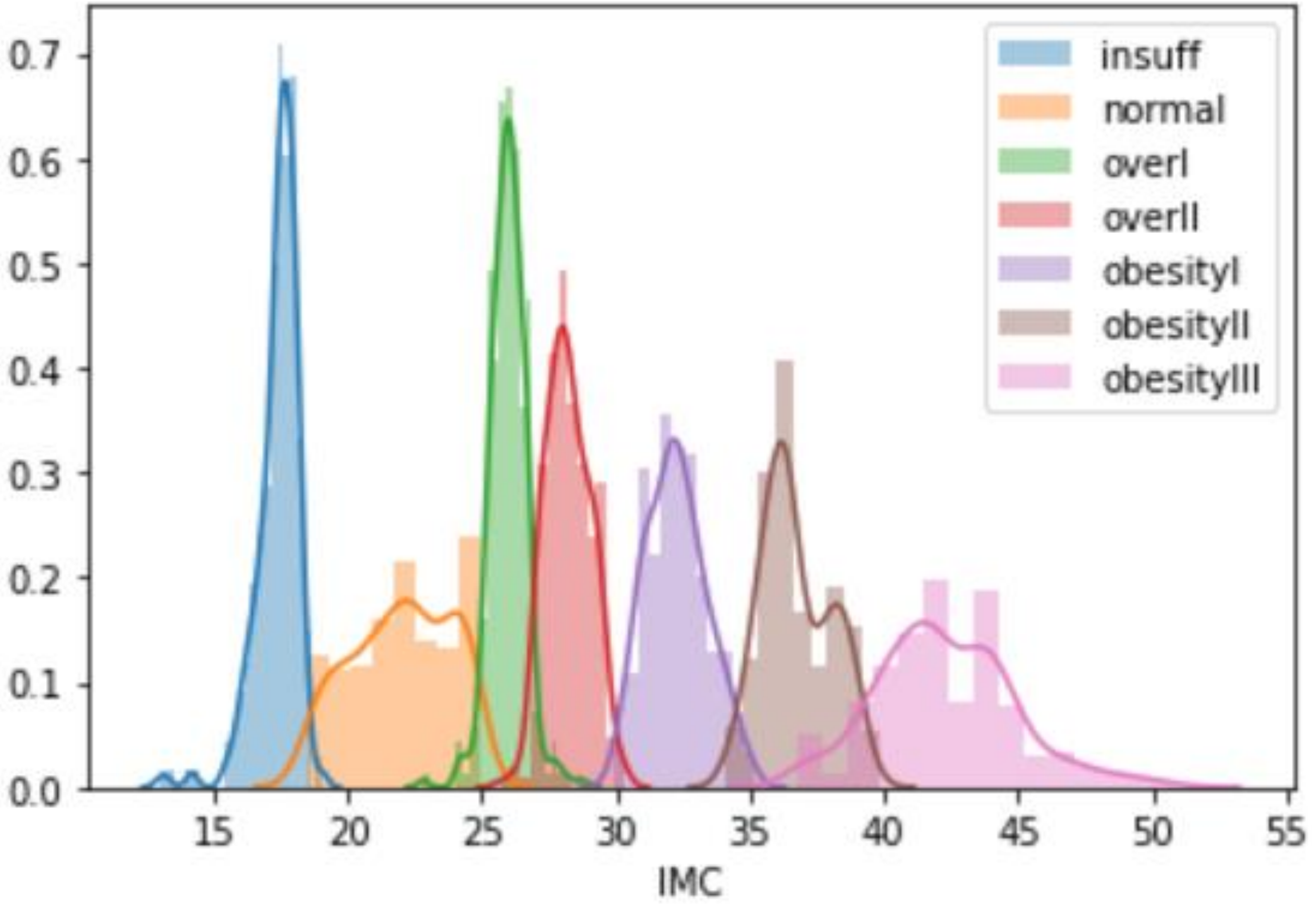
# The Dataset

- 17 columns
- 2111 lines
- 8 quantitative variables
- 9 qualitative variables

There is no N.A value in the dataset !

# Selected variables

- We have studied the relations between variables and the target variable
- Height and Weight seemed to be important in the model
- So, we created the IMC value (Weight/Height²)



The results are excellent, but, is it what we realy want ?

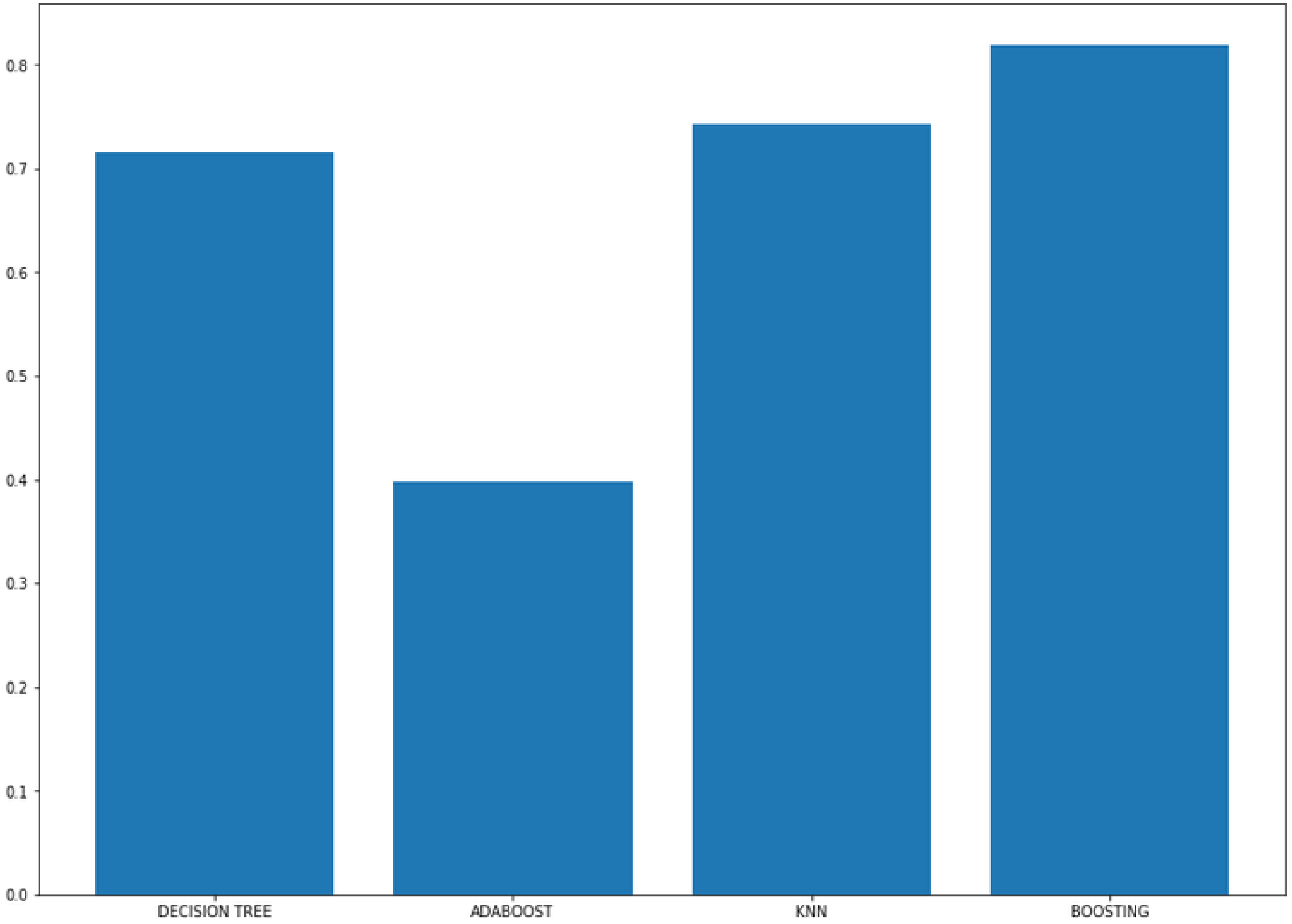
# The New Approach

- For this new step, no more Weight and Height (or IMC) variables

- We have tried 4 Machine Learning model :

  - Decision Tree
  - ADA Boost
  - Knn
  - Boosting

- Boosting look to be the best model



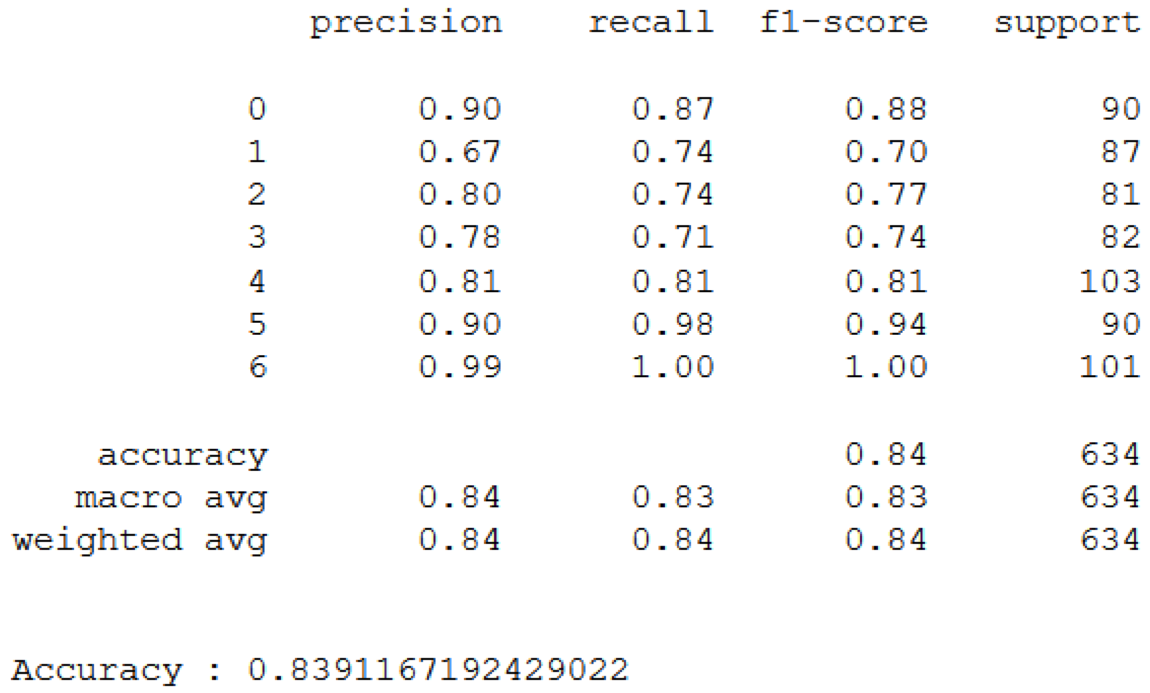
Accuracy for each model

# The Results

After a GridSearchCV, we found the best parameters for this model.

We reached an accuracy of 0,839 !

(Without using Height and Weight)



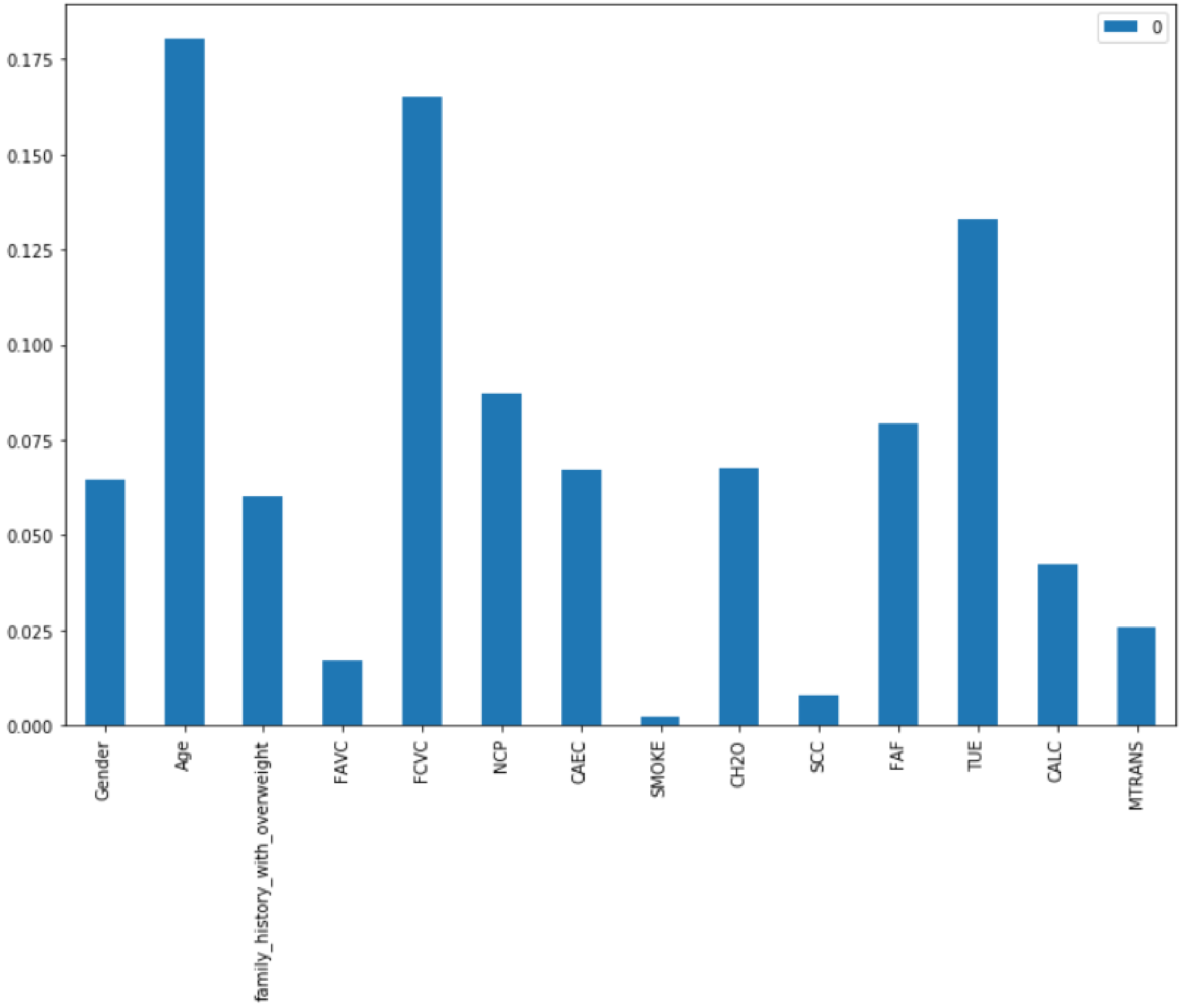
```
              precision    recall  f1-score   support

           0       0.90      0.87      0.88        90
           1       0.67      0.74      0.70        87
           2       0.80      0.74      0.77        81
           3       0.78      0.71      0.74        82
           4       0.81      0.81      0.81       103
           5       0.90      0.98      0.94        90
           6       0.99      1.00      1.00       101

    accuracy                           0.84       634
   macro avg       0.84      0.83      0.83       634
weighted avg       0.84      0.84      0.84       634


Accuracy : 0.8391167192429022
```

# Conclusion

- This study was very interesting. Indeed, obesity in Mexico, Columbia and Peru is a real issue of our time. The results show the impact of eating habits and living behaviors.

- Also, your age, how often you eat vegetables, and how much time do you spend on screens and technological devices are major factors for health.

## Importance of each variables


Gender
Age
family_history_with_overweight
FAVC
FCVC
NCP
CAEC
SMOKE
CH2O
SCC
FAF
TUE
CALC
MTRANS