Estimation of obesity levels based on eating habits and physical condition

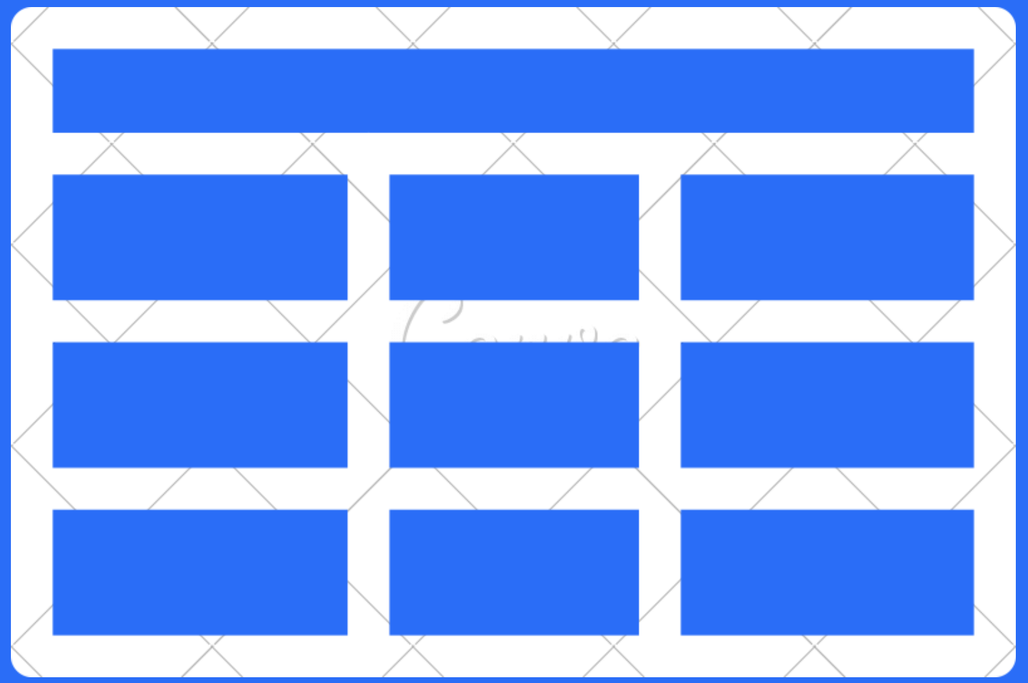
NADESU Pratheswar
NENEZ Owen

Dataset

Dataset of individuals from South America



17 attributes
2111 records

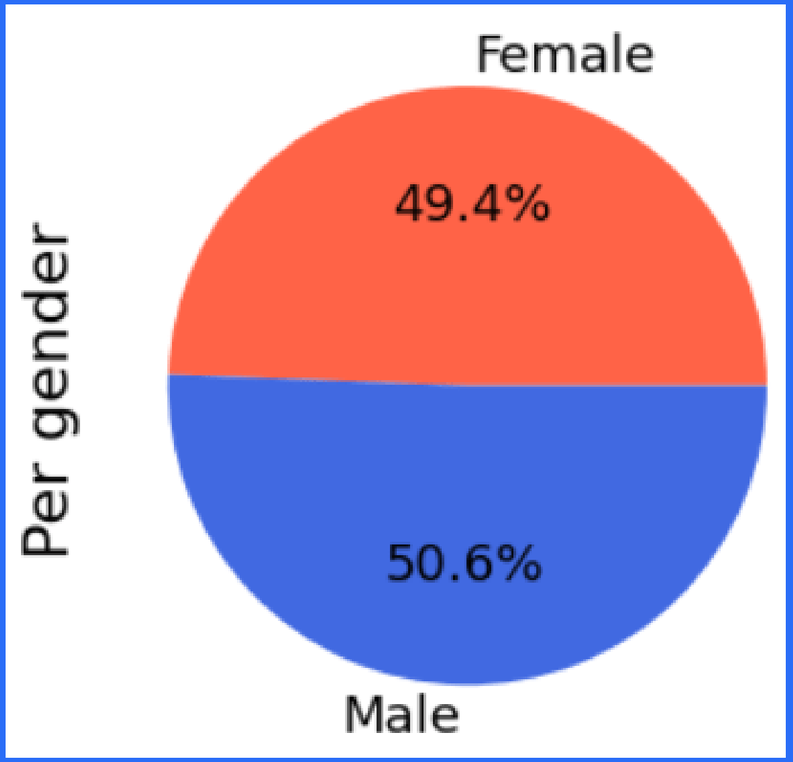


NObeysdad is our target variable

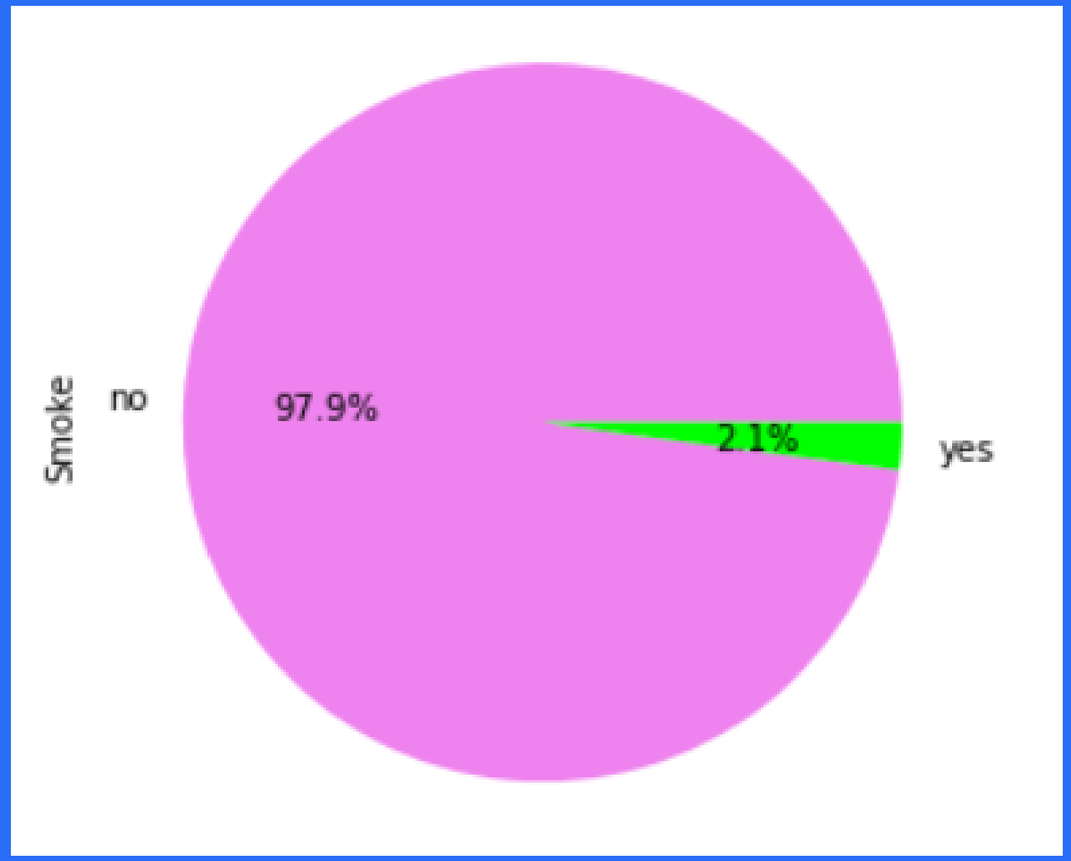
NObeysdad is a *qualitative variable* which allow to see for each individual whether they are in overweight or not.

A dataset evenly balanced

We have almost the same number of men and women in the dataset with makes our analysis more **unbiaised** by the gender factor.



A non smoker population

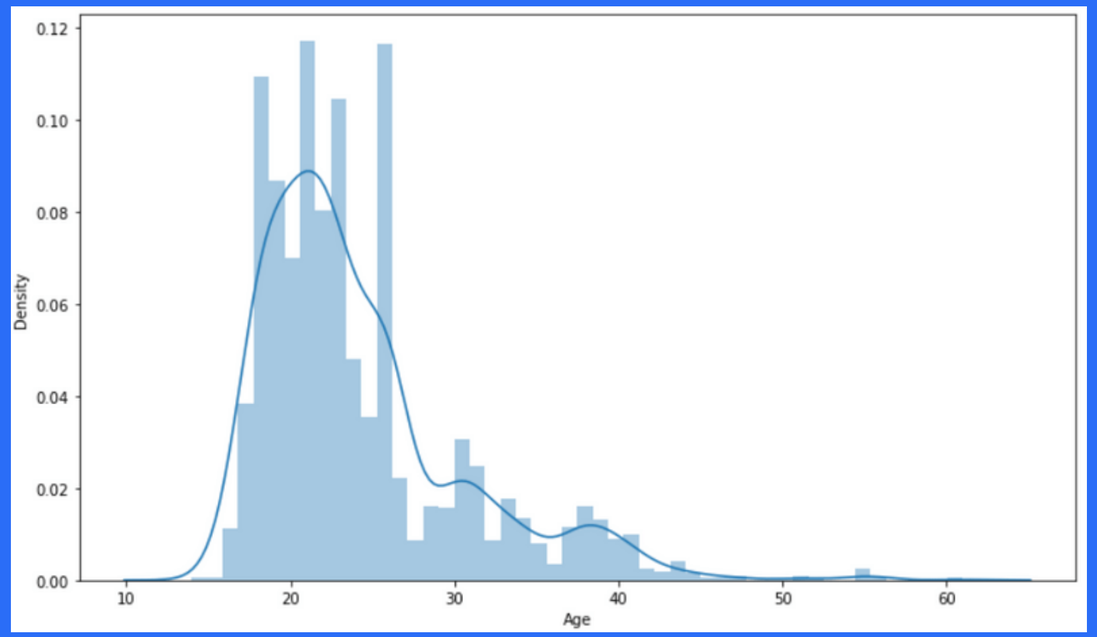


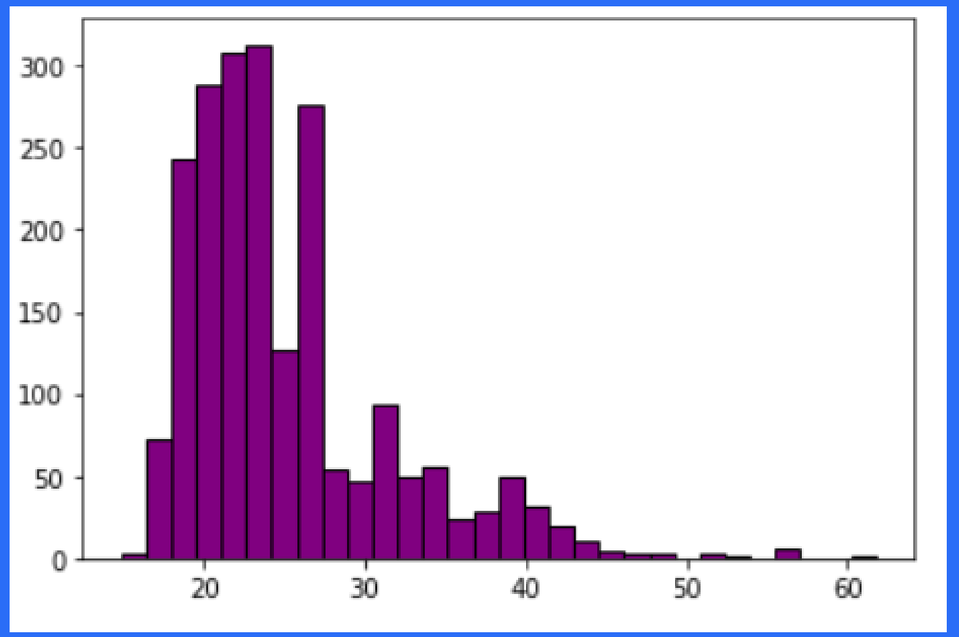
There is a large majority of non smoker individuals. It is said that smoking causes obesity. As a result, this data can help us see what other variable would cause obesity and thus potential cardiovascular risk.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4401671/

A rather younger population

According to these density graph, the population is around 20-30 years old. Thus, we can find **relevant** obesity cases as humans tend to grow fatter as they age.



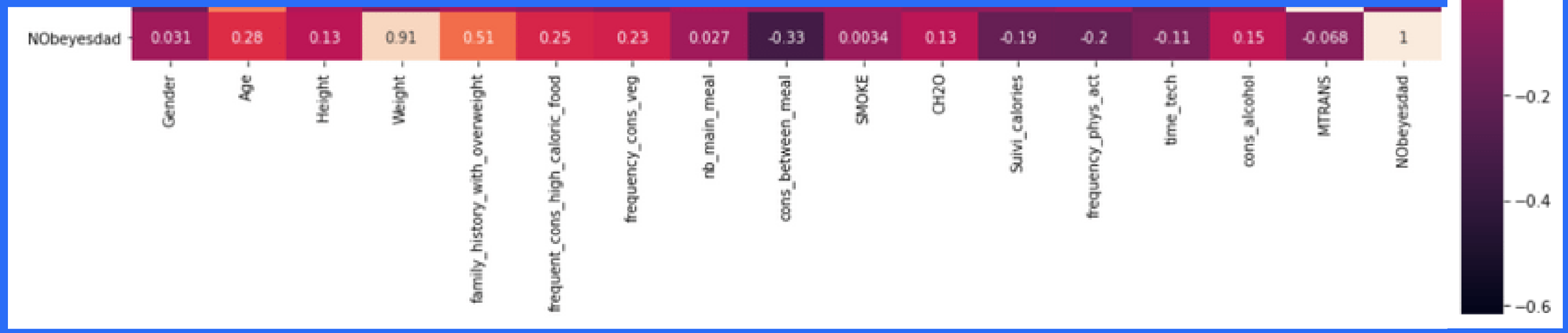


https://www.npr.org/templates/story/story.php?storyId=123887823&t=1641424966398

Correlation between variables

We see that there is a high correlation between our **target variable** and the **weight**.

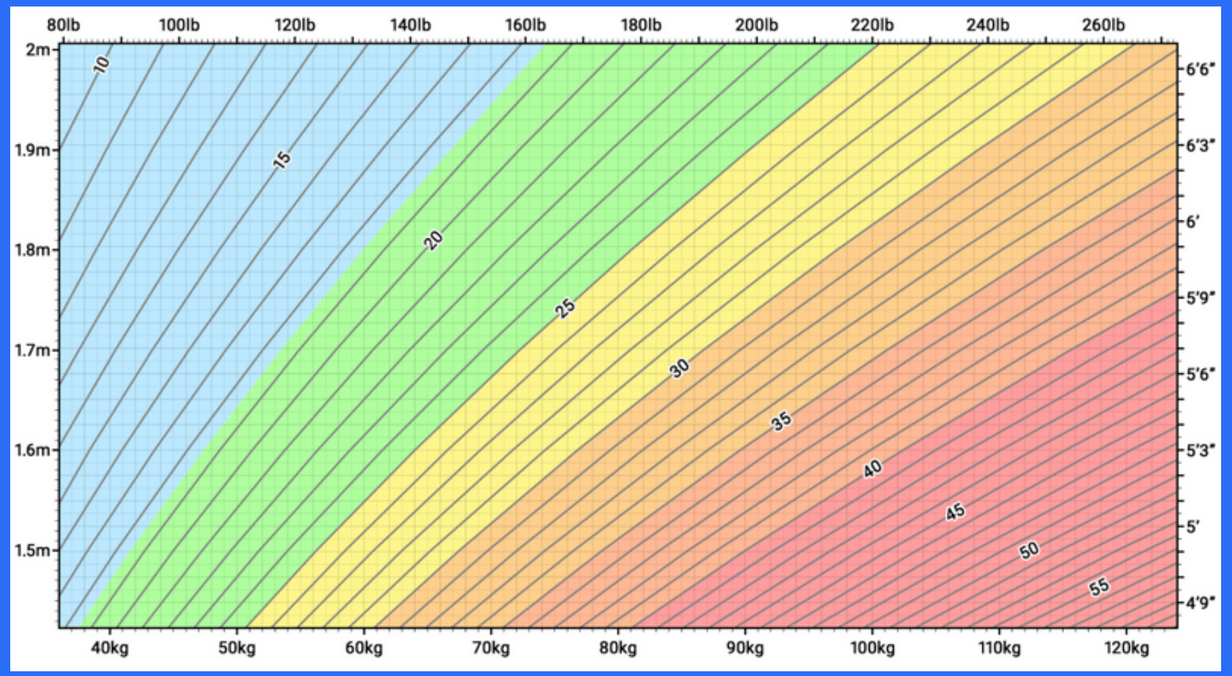
There are less correlation between our **target variable**, the **age**, the **height**, the **frequent consommation of high caloric food** and **family history with overweight**.



Correlation between height, weight and obesity levels

Obesity levels can be explained with the ratio between height and weight..
In fact, obesity levels are calculated with **BMI (*Body Mass Index*)** :

$$\mathrm{BMI} = \frac{\mathrm{mass_{kg}}}{\mathrm{height_m}^2}$$



Machine Learning

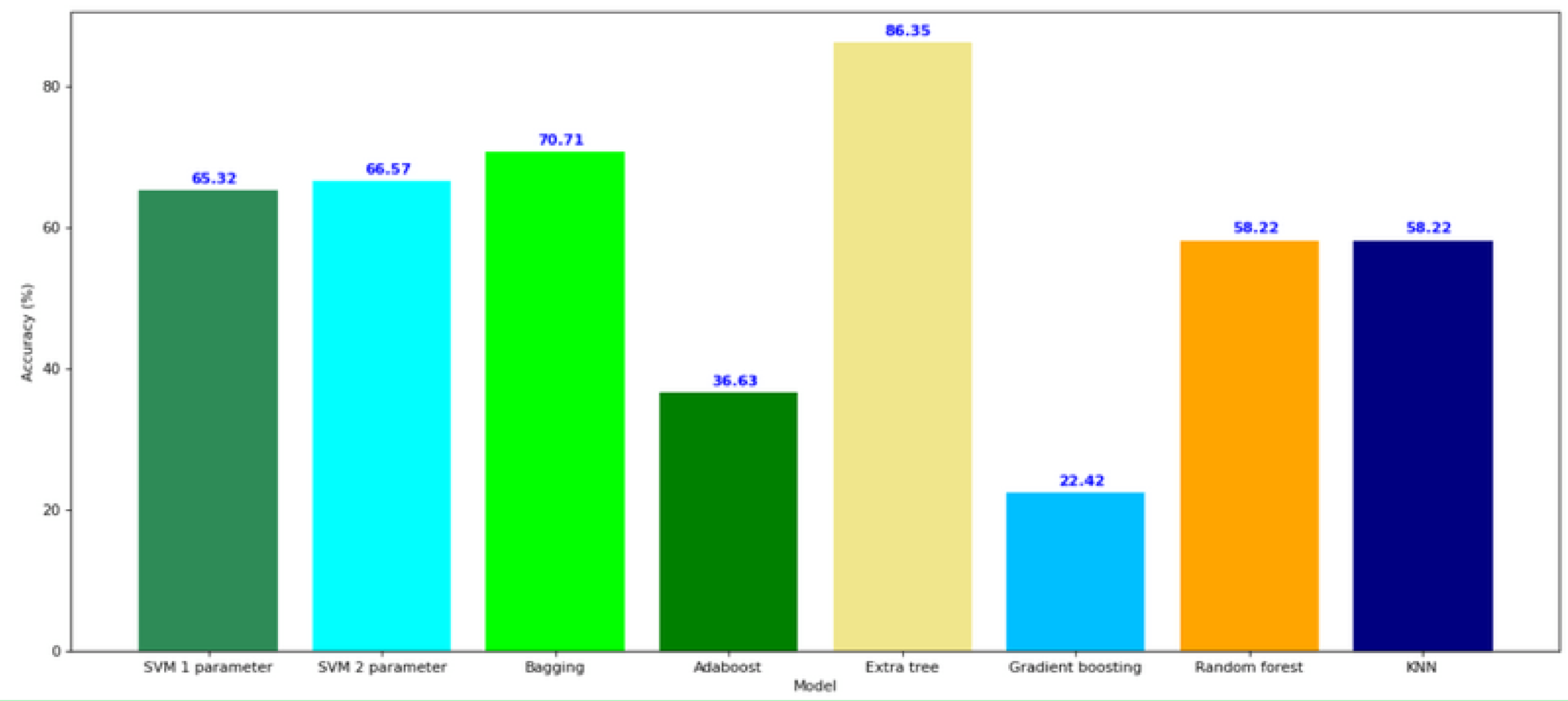
In order to make our prediction on obesity levels, we decided to remove weight and height variable because they are highly correlated to our target variable.

We want to see the other potential causes of obesity.

8 models

Adaboost
Random Forest
Extra tree
Gradient Boosting
SVM (2 models)
KNN
Bagging

Scores of different models



Score is the ratio of relevant predictions.

Conclusion

The best model is **Extra Tree**.

Apart from weight and height,
family_history_with_overweight,
frequent_cons_high_caloric_food,
age are potential factors to obesity and cardiovascular in South America.