# How our lifestyle can cause obesity

12 GIUGNO

**1053015: Nicola Zambelli**
**1057874: Mohammed Amine Bikhtancer**

UNIVERSITAS STUDIORUM BERGOMENSIS

# Abstract

Obesity is now officially identified as a global epidemic, considered one of the greatest health problems of the 21st century. Since 1975 the worldwide prevalence of obesity has almost tripled and stood at 13%, with an absolute value of 650 million obese individuals over the age of 18(WHO, 2022). The following report has the purpose to analyse the results of a survey on obesity to find the correlation between a specific lifestyle and the level of obesity a person can be diagnosed to. With the use of tree-based method we will identify the best model to predict if a person is obese. The logistic regression allows us to make observations on which are the most influential variables to impact on the obesity index.

# Body

## Dataset description

The dataset is provided by the Universidad de la Costa (in Colombia) that got the data through a survey and added some records through the Weka's filter SMOTE. The survey asked the attendants some specific questions to valorise 17 attributes:

1. Related to eating habits:

    - Frequent consumption of high caloric food (FCCI): binary variable('No'=0,'Yes'=1)
    - Frequency of consumption of vegetables (FCV): discrete variable
    - Number of main meals (NP): continuous variable
    - Consumption of food between meals (MFP): discrete variable (From 'Never'=0 to 'Always'=3)
    - Consumption of water daily (CWater): continuous variable
    - Consumption of alcohol (CALC): continuous variable

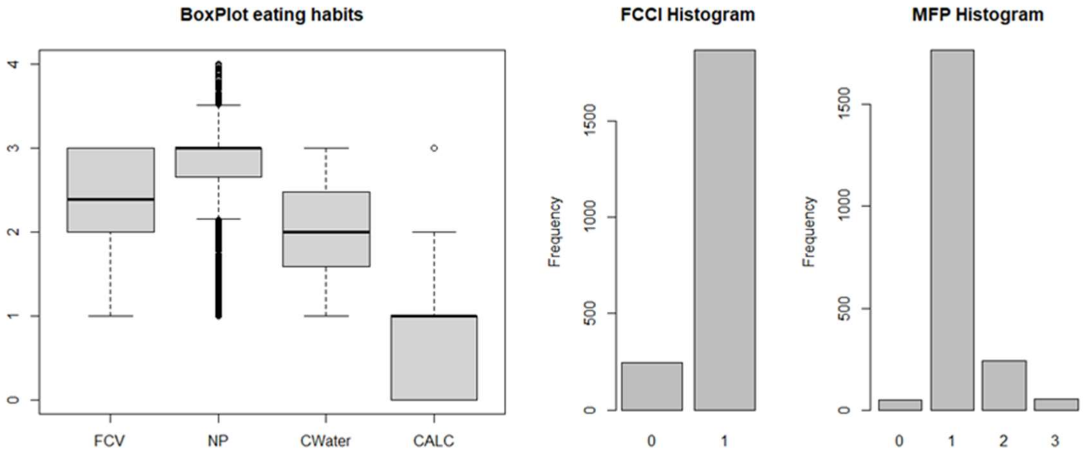

*Figure 1: eating habits boxpot and histogram*

2. Related to physical conditions:
    - Calories consumption monitoring (SCC): binary variable
    - Physical activity frequency (FAF): continuous variable
    - Time using technology devices (TUE): continuous variable
    - Transportation used (MTRANS): not orderable discrete variable. In this case we opted for the "one-hot-encoding" technique, so we created a binary variable for each category.
3. Other generics
    - Gender: binary variable('Female'=0,'Male'=1)
    - Age: discrete variable
    - Height: continuous variable
    - Weight: continuous variable
    - Smoke: binary variable
    - The family history of overweight condition: binary variable

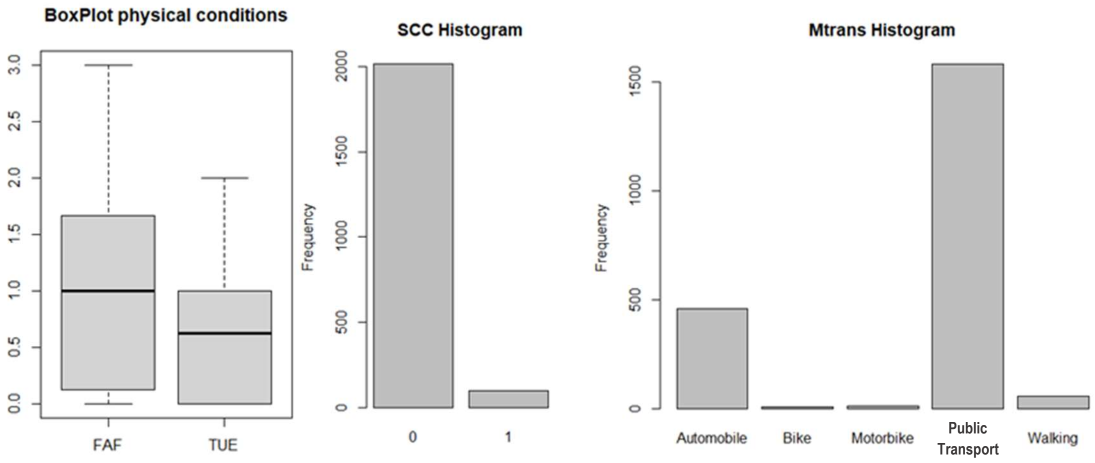

*Figure 2: physical condition boxplot and Histogram*

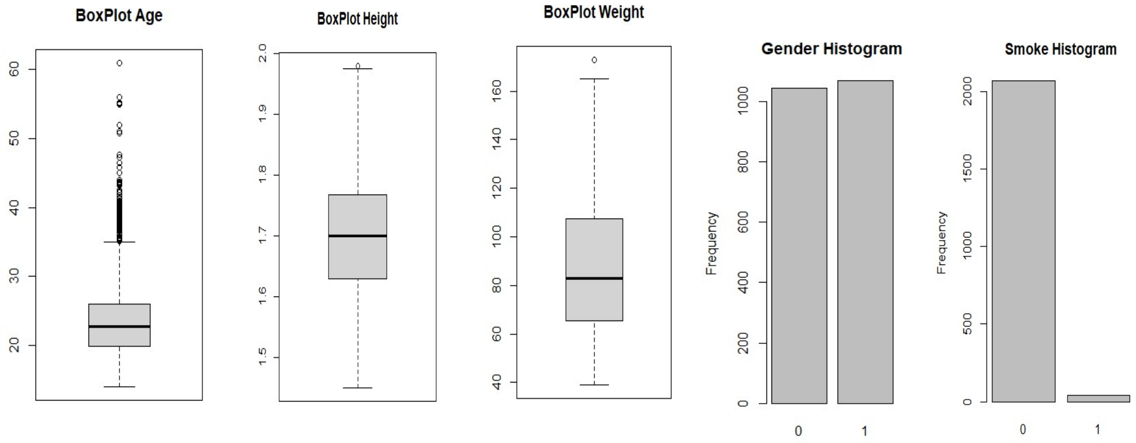

*Figure 3: generics boxplot and Histogram*

The body mass index (BMI) is a measure that uses your height and weight to work out if your weight is healthy.

$$BMI = \frac{Weight}{Height^2}$$

The classes of obesity are divided by BMI into:

- Underweight Less than 18.5
- Normal 18.5 to 24.9
- Overweight 25.0 to 29.9
- Obesity I 30.0 to 34.9
- Obesity II 35.0 to 39.9
- Obesity III Higher than 40

Finally, we thought it was useful to add a column in the dataset that flagged if the person is obese or not, or that it had a BMI greater than 30.

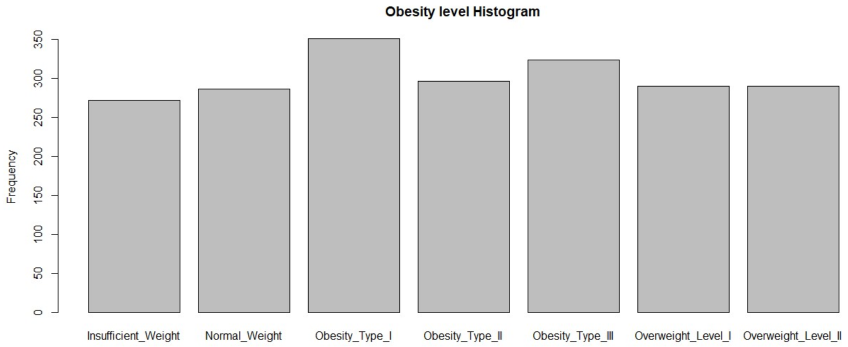

*Figure 4: Frequency of obesity*

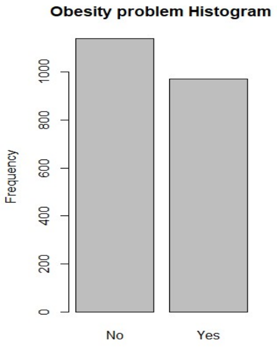

*Figure 5: Numbers of obese people*

Nearly half of the record suffers from obesity.

# Scientific questions

- Do eating habits influence the condition of obesity?
- Do physical habits have implications for the obesity condition?
- Which ones are more influential?

# Methodology

There is a mathematical relationship between obesity (BMI) and weight and height, so, we will not consider these two variables when making statistical inference.
Since our dataset was dealing with categorical variables and since the problem, we are dealing with is the classification of a person as affected by obesity, we started to analyse the data through the classification trees, using the *gini* minimization function.
A complex tree can be difficult to interpret, we use *cross-validation* with different levels of tree size, to understand if it is better to take a simpler tree, trying not to increase the test-error too much.
To improve the tree-based models, we resort to *bagging* with 350 trees and successively we use *random forest*, looking for the best value of *m* predictors chosen as split candidate from the full set of *p* predictors. the goal is to find the methodology that best fit our data at the cost of losing interpretability. We can however obtain an overall summary of the importance of each predictor using RSS. For the same purpose, we use boosting for compare the results.
Afterwards we proceeded to use the *logistic regression* to understand the impact of each regressor on the decision "Which is the probability that this person is obese?". With logistic regression we find models that lead to a test error greater than that made by tree-based methods, but in the other hands they are more reliable in drawing conclusions on the influence of a regressor. In looking for the best logistic regression model we started with the complete least squares model containing all predictors *p*, then iteratively removes the less useful predictor, one at a time, according to the *Backward Stepwise Selection*. Finally, we use cross-validation to find the optimal classification threshold.
To evaluate the performance of the models found, we will often resort to the *confusion matrix*, from which we calculate the performance of *sensitivity*, probability that an obese is positive in the model, and *specificity*, probability that a non-obese is negative.

## Data analysis

The regression trees were used to understand how the regressors affects the obesity condition, at first, we built a tree just for eating habits, then for physical habits and finally one tree based on both the sets of regressors. First, we divided the dataset into two sub datasets, one for the train and one for the test phase. at the first iteration we have obtained the best decision tree (fig.6 top left), but it is easy to notice that despite having good performances it is very complicated. therefore, through the cross-validation technique we want to look for a tree that is less complex but maintains good characteristics. the result is the graph that relates the misclassification error and the size of the tree, in terms of the number of nodes. (fig.6, bottom left) In the case of the eating habits tree, taking a tree of 5 nodes, a less complex tree is obtained with a good performance compromise. the pruned tree obtained is easy to interpret (fig.6, right).

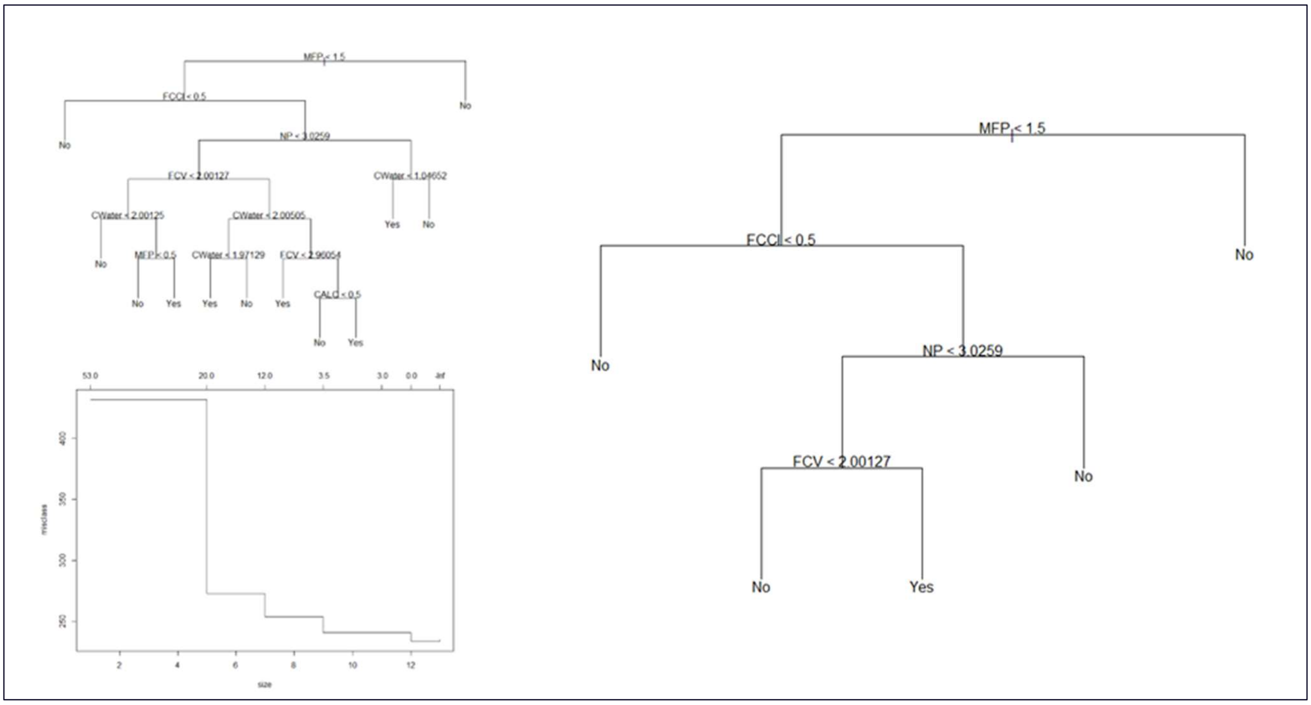

*Figure 6: the eating habits decision tree*

to understand how much precision, we are willing to lose to increase interpretability, we show the confusion matrices and performance of the original tree and the pruned tree evaluated un the test dataset.

| Predicted value | NO | YES |
|---|---|---|
| NO | 436 | 150 |
| YES | 127 | 343 |
| Performance analysis | | |
| Sensitivity | 77,44% | |
| Specificity | 69,57% | |

| Predicted value | NO | YES |
|---|---|---|
| NO | 434 | 95 |
| YES | 129 | 398 |
| Performance analysis | | |
| Sensitivity | 77,09% | |
| Specificity | 80,07% | |

*Table 1&2: eating habits confusion matrix and performance, on the left the pruned tree, on the right the original tree*

The original tree has a test error rate around 21%, the pruned tree has a test error rate of 26%.

The same decision-making process was conducted considering physical habits. (fig.7) Also in this case, we have opted to prune the tree until we get five knots. The difference between the test errors committed by the original tree and the pruned tree is minimal, in fact we have respectively 33.1% and 33.3%. in general, this model is worse than the previous one.

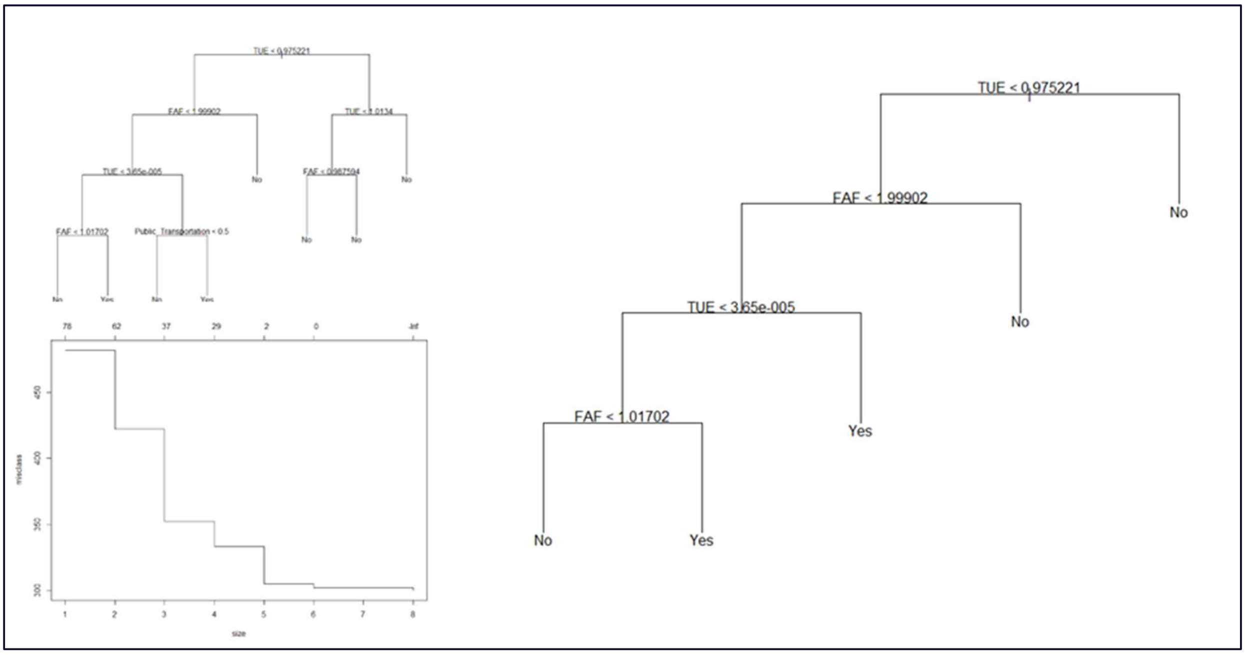

*Figure 7: the physic habits decision tree*

Finally, we perform the same operations including all the decision variables. in this case we decide to get a pruned tree with 6 nodes. although we have included more variables, we are unable to obtain a tree that gives us a test error rate of less than 20%. The trees obtained are unbalanced this could be a problem due to the starting dataset, in fact many categorical variables are concentrated on a few classes. To improve the model, we need more powerful methodologies such as bagging or random forest.

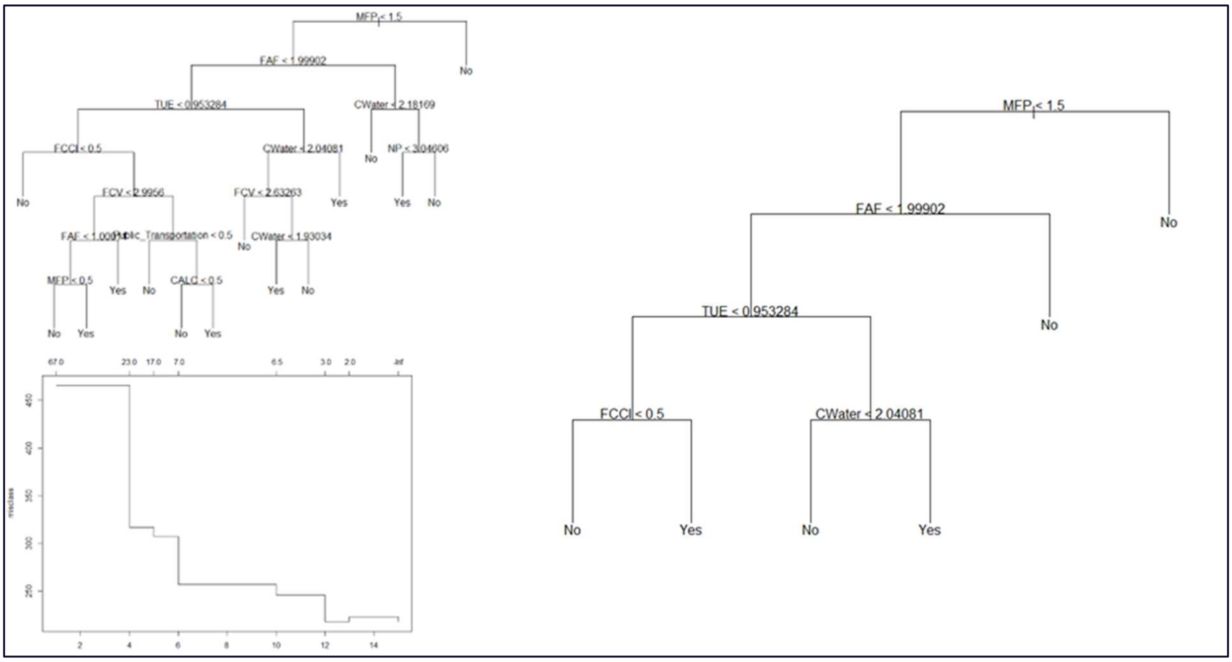

*Figure 8: overall decision tree*

By applying bagging with 350 trees, the model improves considerably to a test error of 12.8%. then, we use the random forest technique by choosing *m* equal to the nearest integer of the root of *p*. in this case we get a slightly better model considering the train error, but we have the same test error obtained before.

We wonder if there is an *m* that allows us to obtain a model with better performance. thus, we evaluate how train and test error vary as *m* varies(fig.9).

The result shows that by choosing m equal to 5 we can obtain the best model, with a test error of 11.6%.

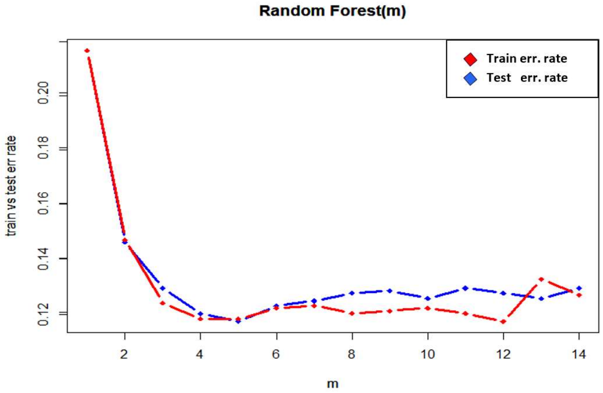

*Figure 9: Random Forest with different values of m*

Another confirmation of the fact that, in our case, random forest is better than bagging, is given by the graph of the error committed when the trees considered grow(fig.10). the blue line is given by random forest with *m* equals to five and the red represents bagging. it is easy to see how the blue line often stays below the red one

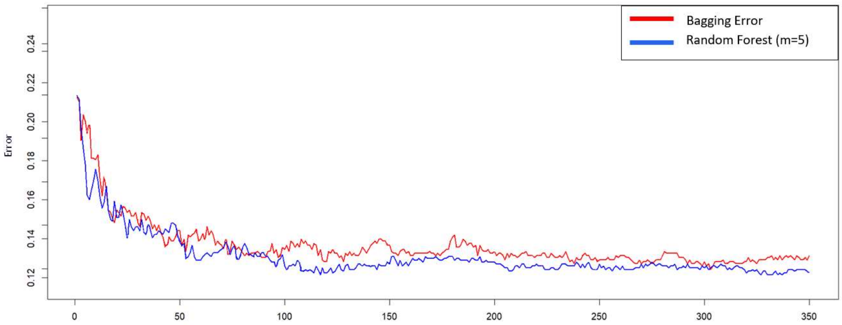

*Figure 10: Bagging VS Random Forest(5) Error*

Once we have found the model that satisfies us, we are interested to understanding the importance of decision-making variables. we apply the boosting technique and compare which are the most influential variables of the two techniques.
With bagging we noticed some expected results, through the importance command we found that some variables perform better than others. We note how the dummy variables obtained from the decomposition of the MTRAS variable give us little information. The problem could derive from the fact that some of these variables have few observations: the variable bike for example, has only 7 records, i.e., only seven people out of more than 2000 use the bike as their preferred means of transport.

|  | No | Yes | MeanDecreaseAccuracy | MeanDecreaseGini |
|---|---|---|---|---|
| FCCI | 12.892984 | 39.655122 | 38.904595 | 31.00882203 |
| NP | 27.383060 | 54.167005 | 54.921486 | 84.33819439 |
| CWater | 22.590003 | 32.872893 | 36.684183 | 80.25670441 |
| FCV | 22.766778 | 49.923056 | 52.358066 | 72.53795764 |
| CALC | 8.016679 | 24.922064 | 25.100283 | 26.17723705 |
| MFP | 12.861284 | 64.940620 | 52.209516 | 57.35185170 |
| SCC | 12.580189 | 24.106986 | 24.505594 | 11.27581933 |
| FAF | 14.020705 | 33.831751 | 33.505951 | 64.80297893 |
| TUE | 22.063509 | 32.737071 | 37.850421 | 63.08115480 |
| Public_Transportation | 7.596204 | 20.978616 | 21.377804 | 14.03542427 |
| Walking | 1.694699 | 7.371395 | 7.241769 | 2.15484453 |
| Automobile | 4.920759 | 15.463164 | 15.639503 | 10.17390179 |
| Motorbike | -1.034697 | 0.000000 | -1.081562 | 0.53390532 |
| Bike | 0.000000 | 0.000000 | 0.000000 | 0.04135992 |

*Table 3: Relative influence detected through bagging*

NP and MFP, i.e., the consumption of vegetables and having more meals than three, are very influential for the condition of obesity, how confirmed by the results of boosting:

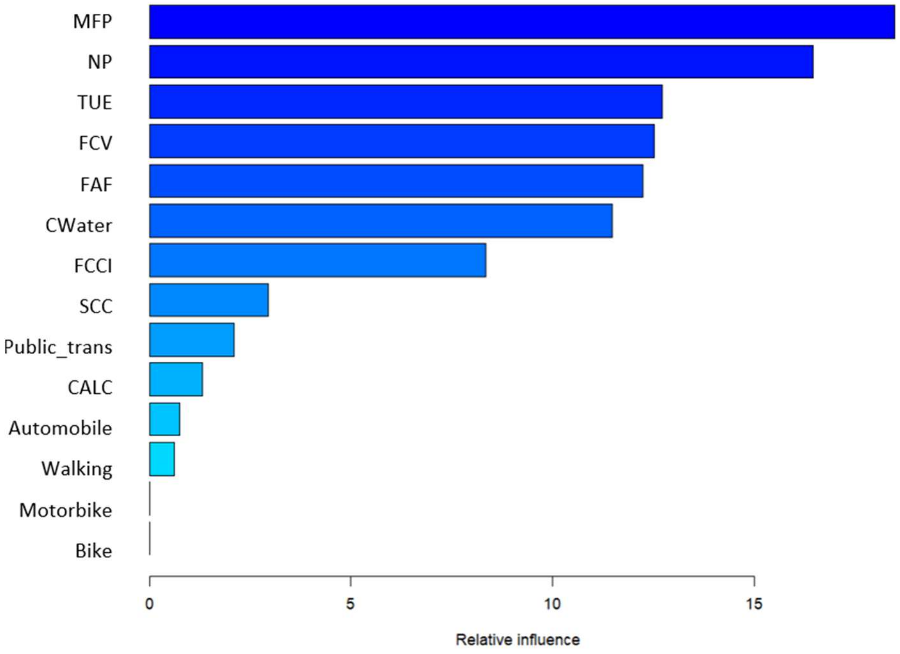

*Figure 11: Relative influence detected through boosting*

Since the trees suffer from high variance, since the trees we built had a high RMD we proceed with logistic regression to answer our questions.

Applying the logistic regression with only the eating habits we found out that there are some very useful regressor to predict the output.

| | Estimate | Std. Error | Z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -3.84695 | 0.42615 | -9.027 | <2e-16 | *** |
| FCCI | 2.48562 | 0.25025 | 9.932 | <2e-16 | *** |
| NP | 0.14213 | 0.06336 | 2.243 | 0.0249 | * |
| CWater | 0.16969 | 0.08049 | 2.108 | 0.0350 | * |
| FCV | 0.81200 | 0.09380 | 8.657 | <2e-16 | *** |
| CALC | 0.09078 | 0.09601 | 0.946 | 0.3444 | |
| MFP | -1.22714 | 0.13757 | -8.920 | <2e-16 | *** |

*Table 4: Logistic regression with eating habits*

This model has an error rate of 0.3301753, removing the less useful regressors (NP, CWater and CALC) we got a little improvement since this model has an error rate of 0.3287541.
We then proceeded with the physical habits.

| | Estimate | Std- Error | Z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -0.96156 | 1.10898 | -0.867 | 0.385906 | |
| SCC | -3.30023 | 0.59016 | -5.592 | 2.24e-08 | *** |
| FAF | -0.30124 | 0.05556 | -5.422 | 5.90e-08 | *** |
| TUE | -0.27767 | 0.07821 | -3.551 | 0.000384 | *** |
| Public Transp. | 1.46089 | 1.10686 | 1.320 | 0.186885 | |
| Walking | -1.11937 | 1.25670 | -0.891 | 0.373080 | |
| Automobile | 1.23823 | 1.10962 | 1.116 | 0.264460 | |
| Motorbike | 0.46678 | 1.31487 | 0.355 | 0.722588 | |
| Bike | NA | NA | NA | NA | |

*Table 5: Logistic regression with physical habits*

With these regressors the error rate was 0.4064424, removing the less useful regressors didn’t help: the error actually increased to 0.4258645.
As we did with the trees, we built a model with all the regressors (leaving out height and weight):

| | Estimate | Std- Error | Z value | Pr(>\|z\|) | |
|---|---|---|---|---|---|
| (Intercept) | -8.87293 | 2.07926 | -4.267 | 1.98e-05 | *** |
| FCCI | 2.26159 | 0.27391 | 8.527 | <2e-16 | *** |
| FCV | 0.92127 | 0.11755 | 7.837 | 4.62e-15 | *** |
| CALC | 0.08597 | 0.11136 | 0.772 | 0.440121 | |
| MFP | -1.46829 | 0.18456 | -7.956 | 1.78e-15 | *** |
| SCC | -2.52783 | 0.63852 | -3.959 | 7.53e-05 | *** |
| FAF | -0.26219 | 0.07166 | -3.659 | 0.000253 | *** |
| TUE | -0.16760 | 0.09764 | -1.717 | 0.086068 | . |
| Public Trans. | 0.67038 | 1.98423 | 0.338 | 0.735472 | |
| Walking | -1.71987 | 2.08130 | -0.826 | 0.408609 | |

| | | | | | |
|---|---|---|---|---|---|
| Automobile | -0.68330 | 1.98380 | -0.344 | 0.730516 | |
| Motorbike | 0.94630 | 2.21786 | 0.427 | 0.669618 | |
| Bike | NA | NA | NA | NA | |
| Age | 0.09177 | 0.01260 | 7.284 | 3.24e-13 | *** |
| Obesity$family_history_with_overweight | 4.05210 | 0.37639 | 10.766 | <2e-16 | *** |
| Obesity$SMOKE | 0.52250 | 0.44719 | 1.168 | 0.242645 | |
| Obesity$Gender | 0.14614 | 0.12189 | 1.199 | 0.230551 | |

*Table 6: Logistic regression with both the sets of habits*

With this model we got the smallest error rate: 0.237802; leaving out the less useful regressors we got: 0.2335386.

To improve the performance of the logistic regression model we want to see what happens when the classification threshold (p), by default equal to 0.5, changes. therefore, we evaluated the cross-validation error as the threshold changed, initially from 0.1 to 0.9 (Fig. 12, left). After noting that the optimal threshold is around 0.6, we performed the same technique between 0.55 and 0.65 (Fig. 12, right). The *p* that minimizes the test error is 0.59, which commits a test error of 0.2278, better than the model with the default threshold.

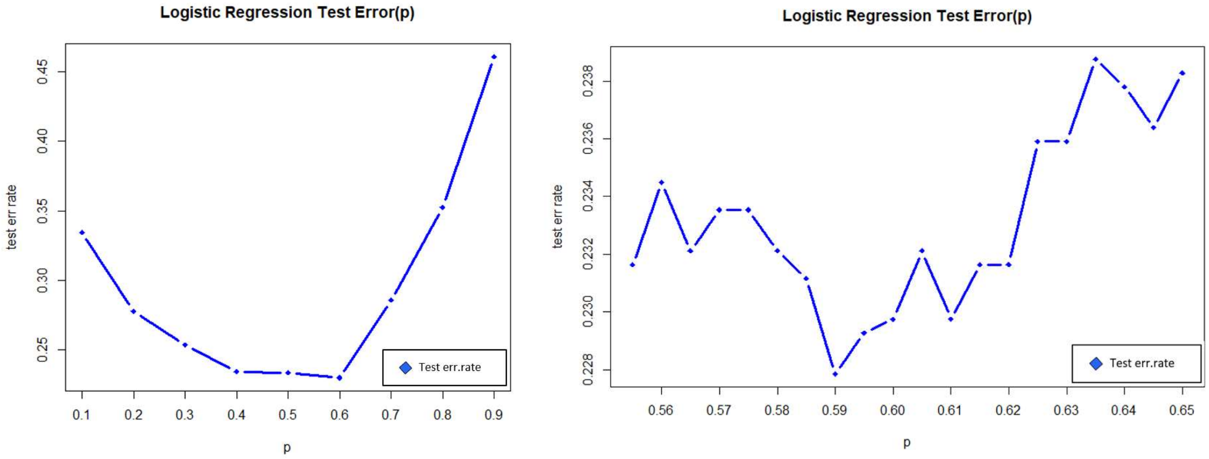

*Figure 12: Logistic regression error rate as the classification threshold changes, [0.1 0.9] to the left and [0.55 0.65] to the right*

| Predicted value | NO | YES |
|---|---|---|
| NO | 927 | 269 |
| YES | 212 | 703 |
| Performance analysis | | |
| Sensitivity | 81,38% | |
| Specificity | 72,32% | |

*Table 7: Logistic regression confusion matrix and performance, classification threshold = 0.59*

# Results discussion

The best model, which makes fewer prediction errors, was found thanks to the random forest technique, choosing an *m* equal to five. In particular, we have shown how this model performs better than the bagging model. We are pleased to have found a model that makes a test error close to 11%, which means that it fails about 1 time in every ten predictions. the analysis carried out using tree-based methods gave some trace on which were the predictors that best describe the problem of obesity. The analysis made, through the logistic regression highlighted the most impactful regressors on the prediction of the obesity condition:

- FCCI (frequency of consumption of high caloric food)
- FCV (frequency of consumption of vegetables)
- MFP (frequency of food between meals)
- SCC (calories consumption monitoring)
- FAF (physical activity frequency)
- Age
- The family history of overweight condition

# Conclusions

We can conclude that considering the eating habits alone, without seeing what lifestyle of a person is conducted is not enough. The best models we got, was that considered both the sets of habits.
In general, methods derived from eating habits are better than those that consider only physical habits. It may mean that a healthy diet is a starting point for preventing obesity problems.
We also determined how important is age on the prediction of obesity, which makes sense since growing older is easier to live a sedentary life (and the frequency of physical activity is one the most important regressors).
Another notable aspect is the importance of having a family member who suffers or has suffered from overweight or obesity problems. it could be that obesity problems could be influenced by genetic factors.
Honestly, an important limitation of our work is the fact that some categories of variables had few observations and consequently were irrelevant, for example people who walk or use the bike as a means of transport.
Moreover, two results are unexpected: the fact that the consumption of vegetables, has a positive effect on obesity and that the bad habit of eating between meals, improves the situation of obesity. This could be a problem due to data sampling, in fact, the sample does not reflect the reality of the population, where the level of obesity, according to the WHO, stood at 13%, not around 40-45%. for example, some data may have been collected during healing process and maybe some patients began to increase vegetable consumption only after diagnosis.
It would be interesting to carry out the same techniques on a more realistic dataset.