10/01/2021

Python For Data Analysis final project

Drugs consumption analysis

Sarah LASCAR & Flora MAIER

Summary

1. Dataset introduction
2. Data cleaning
3. Data visualization
4. Data modelling

1. Dataset introduction

Link of the dataset : UCI Machine Learning Repository: Drug consumption (quantified) Data Set

Subject : The given dataset shows the consumption of 1885 respondents on 19 different drugs, depending on social criterias like personality, age, gender, country or level of education.

Variables

The dataset has 33 variables that we can group :

- Basic informations variables : ID, Age, Gender, Education Level, Country number, Ethnicity number
- Personality level of: Neuroticism, Extraversion, Openness to experience, Agreeableness, Conscientiousness, Impulsiveness, Sensation seeing
- Consumption level of : Alcohol, Amphet, Amyl, Benzos, Caffein, Cannabis, Chocolate, Coke, Crack, Ecstasy, Heroin, Ketamine, Legal highs consumptions, LSD, Meth, Mushrooms, Nicotine, Semeron, VSA

2. Data cleaning

- We checked that all values are available : there is no NA values in the dataset

-We deleted the respondents that declared using “Semeron” because it is a fictive drug we can conclude that they are liars and we can’t trust them.

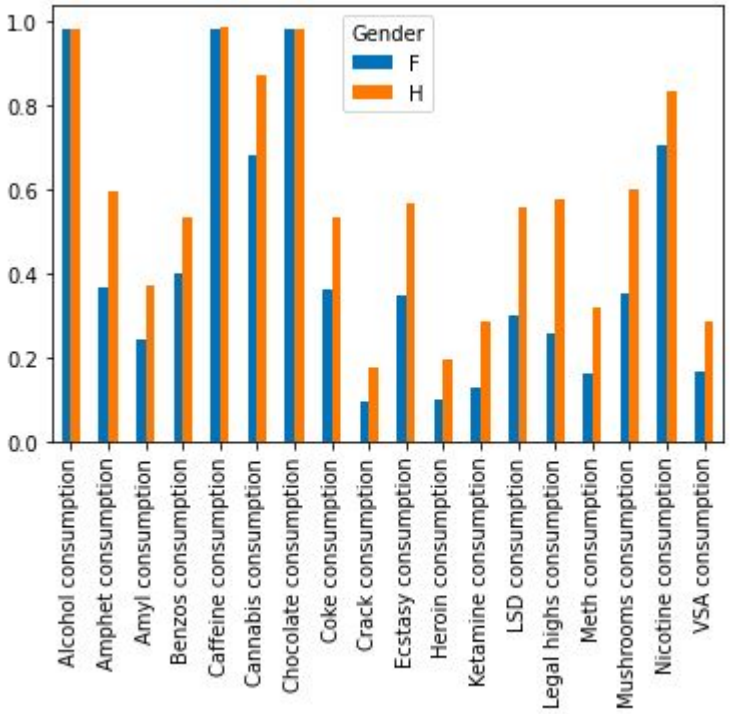
- We converted all the numeric values of “Basic informations variables “ into the string categories they refer to. For example, the gender went to “Male” instead of the numerical value -0.48246.

3. Data Visualization

As the percentages for each column are already provided by the database website, there is no need to use graphs to visualize the profile of respondents in this study.

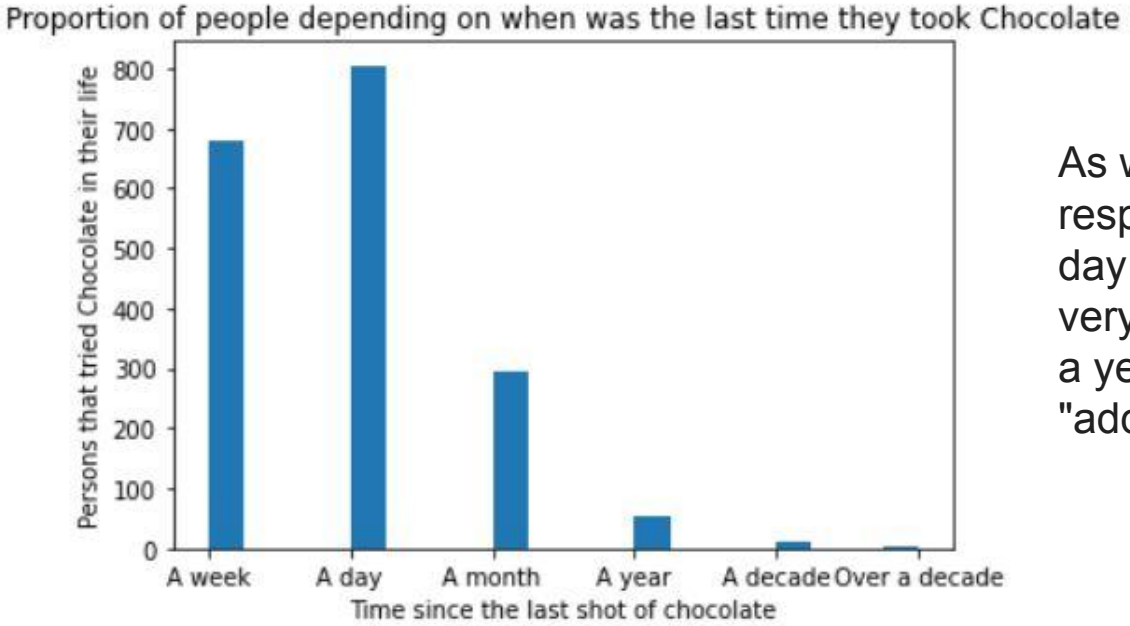
They are mostly white people (both women and men) from the USA or UK, who are between 18 and 54 years old, most of them entered the university.

The consumption of each drug by gender



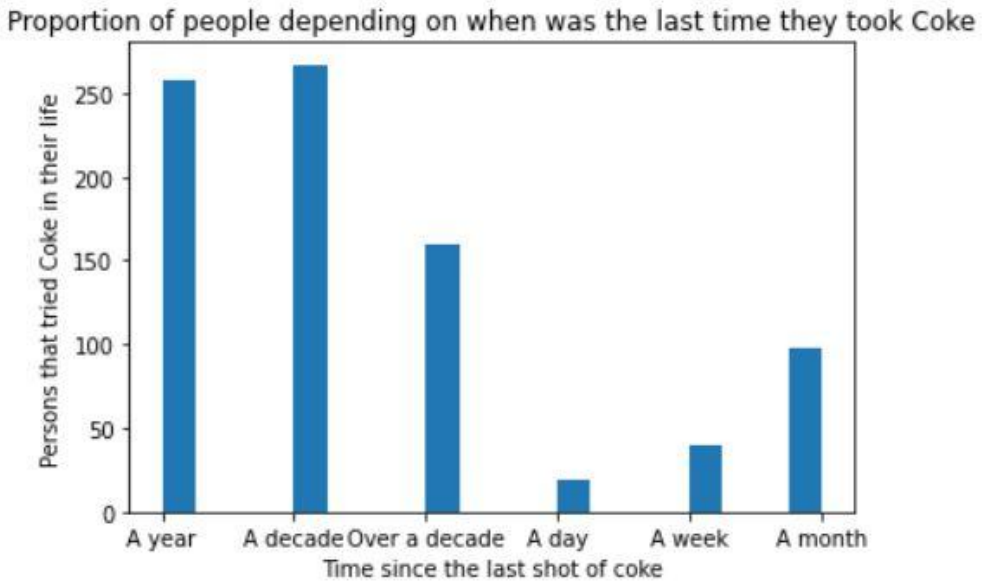
We can clearly see that men are more into drugs than women, except for legal ones when they are very similar.

Use frequency for Chocolate



As we see, most of the respondents ate chocolate just a day ago, or a week ago, when very few of them did not eat it over a year. It is very common to be "addicted" to chocolate!

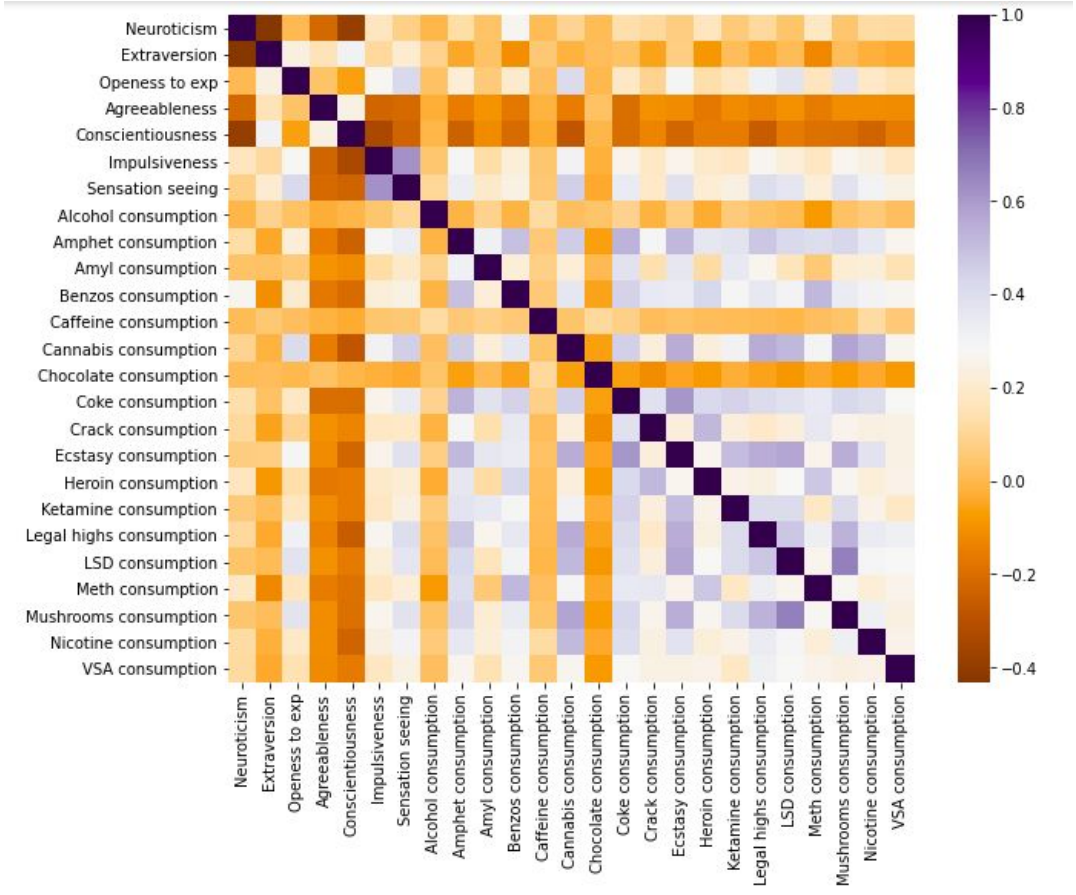
Use frequency for Coke



The results for Coke is totally different that chocolate : most of the persons took coke over a year a go, when only few of them did it a day or a week ago.

We can deduce that people are less addicted to Coke than to chocolate, because Coke is just a passage of their life.

Correlation between drug use and behavior



The consumption of chocolate, alcohol or caffeine is not correlated with the consumption of other substances.

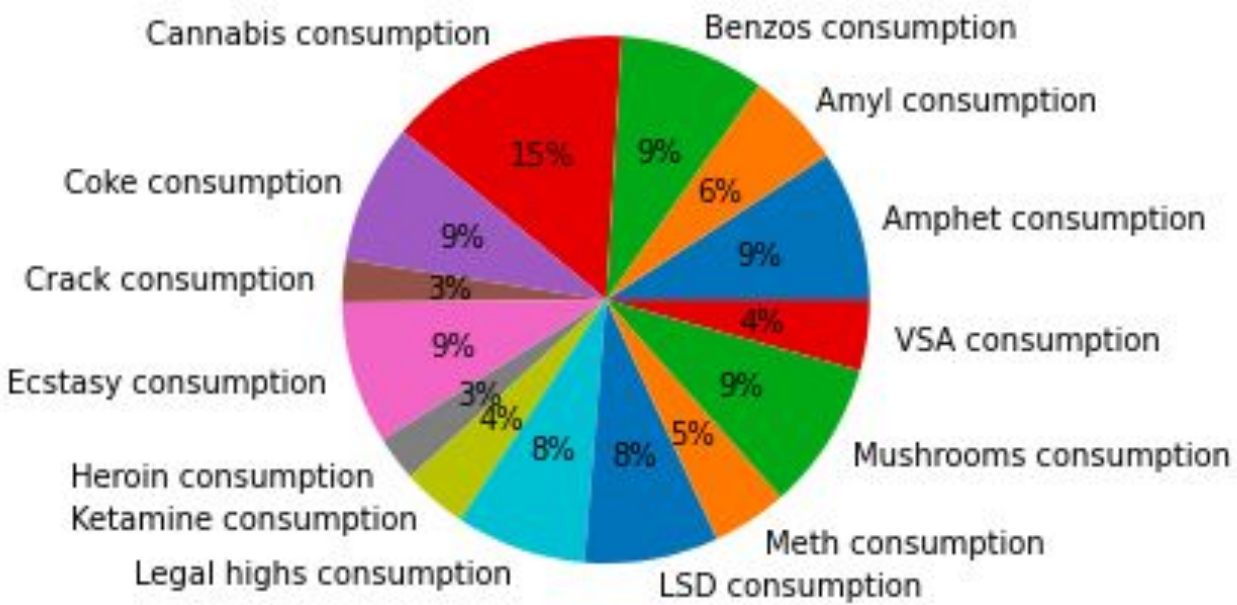
Indeed we will show in the last pie graphics that everyone consume them.

Conversion to a classification problem

From now on, we will split users into two classes: those who have never used the drug in question and those who did it.

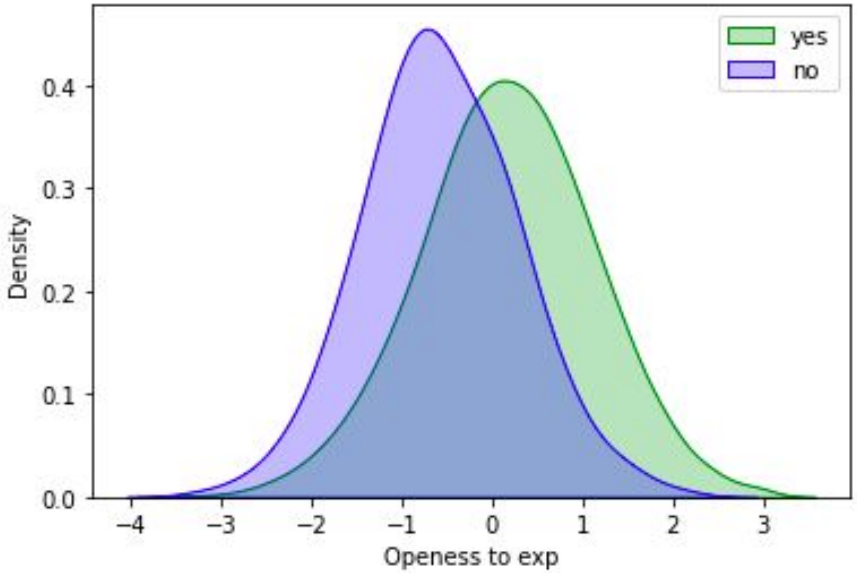
It means that we take in consideration the CL0 value, corresponding to the Non-Use and CL1, CL2, CL3, CL4, CL5, and CL6 all together showing that the person used the drug at least once in their life.

Pie chart of the most used illicit drugs



We can notice that the Cannabis is the most used drug, then Benzos, Amphet, Ecstazy Mushrooms and Coke

Compare the people that already tried Cannabis with the others



The people that already tried Cannabis are more open to experience than the persons that never did it. For the others personalities, it is quite the same.

4. Data Modelling

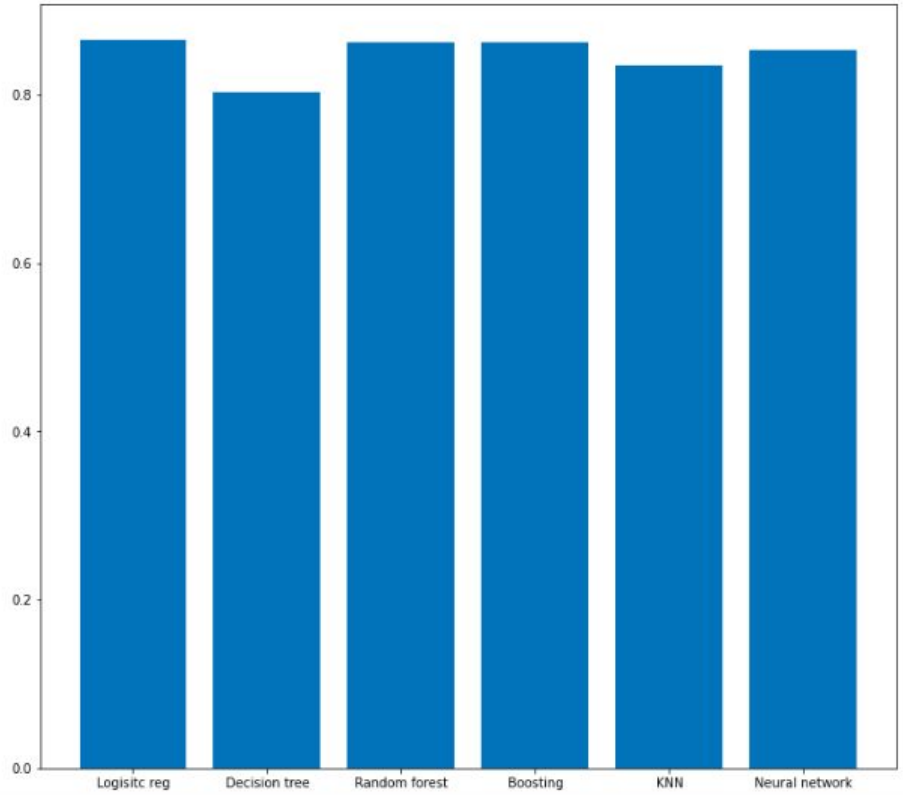
We want to predict if someone already used a chosen drug or not, for all the drugs we saw, depending on his/her sociologic environment and the other drugs he/she is taking.

In order to predict ir, we tested several models:

- Logistic regression model ;
- Decision tree model ;
- random forest model ;
- boosting model ;
- K-Nearest Neighbor model.

Which one is the best model for our dataset?

Average accuracy per model

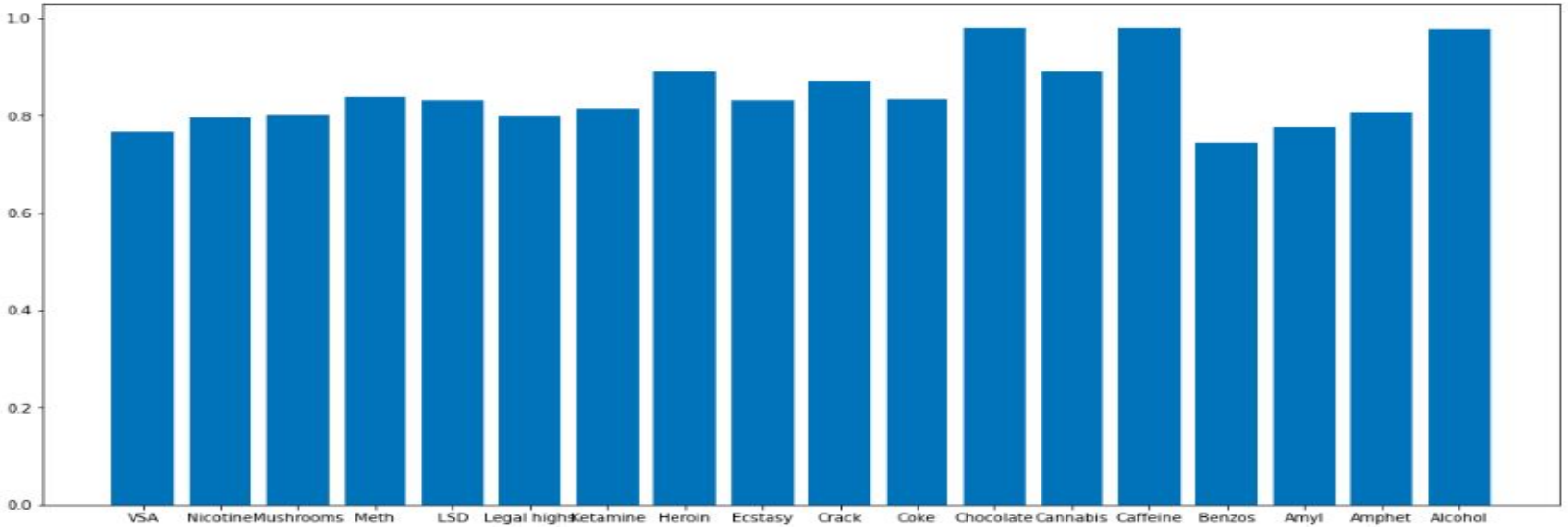


In average we have satisfying accuracies.

Logisitc reg	0.862186
Decision tree	0.810842
Random forest	0.862814
Boosting	0.860215
KNN	0.837993
Neural network	0.853853

The Random forest model is in average the best one, very close to the Logistic regression and the Boosting one.

Average accuracy per model



The easiest drugs to predict are Caffeine, Chocolate and Alcohol, that are by the way the most used one.

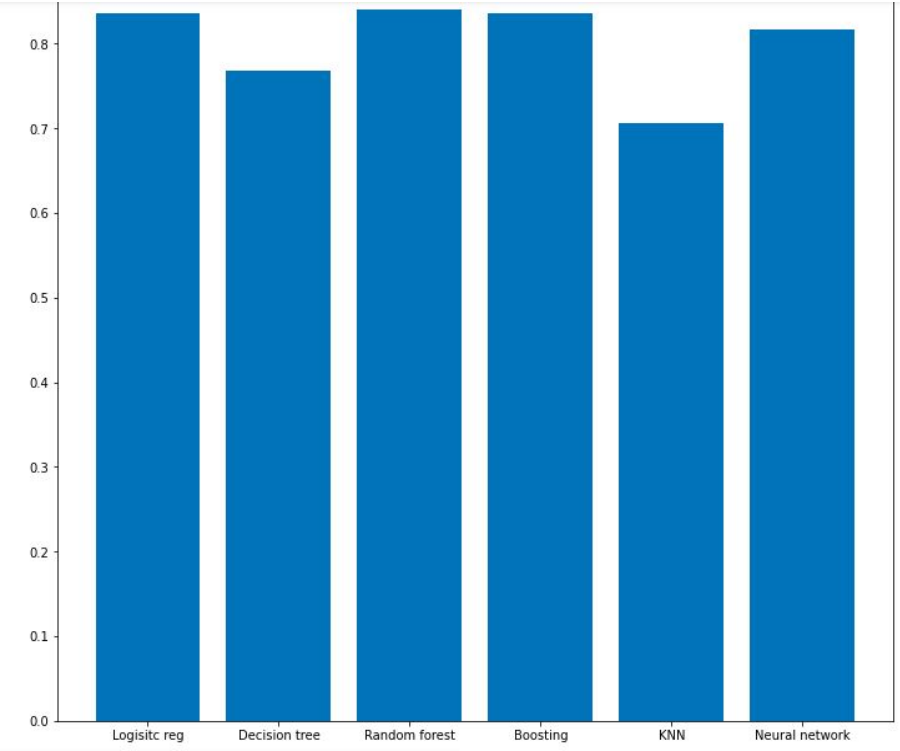
But, after them we can find Heroin and Cannabis and then Crack (Heroin and Crack are the one that are used the less)

Thus we understand than the easiest drugs to predict are the one that are either the most or the least used.

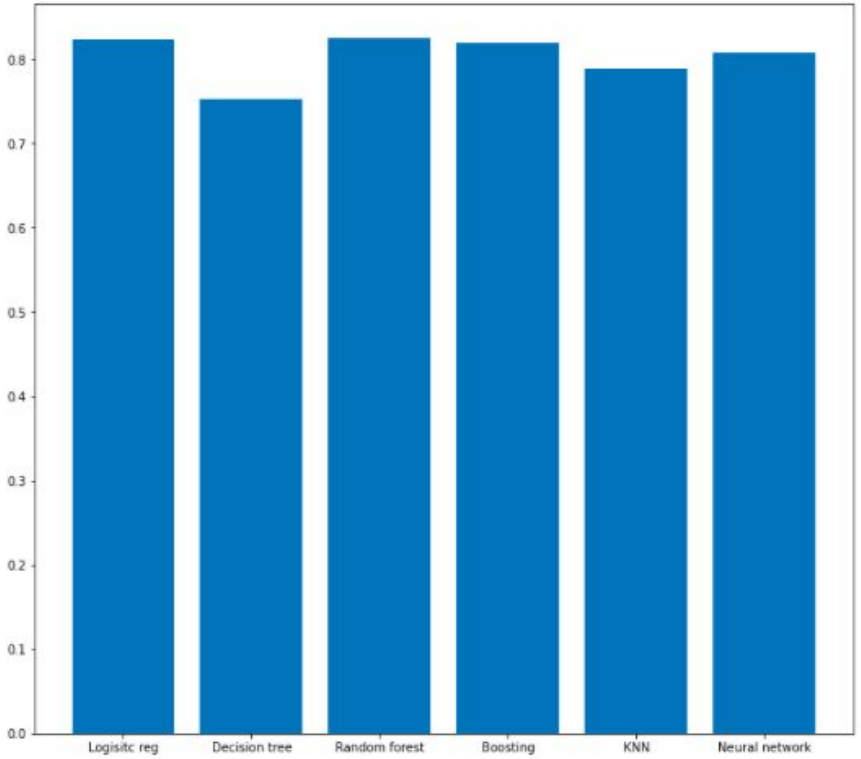
Testing models after some parameters have been removed

As we have seen in the visualization part, Chocolate, Alcohol and Caffeine seemed not very correlated with other drug and neither with behaviors.

We tested the same models as previously without them and tried to predict if someone has already consume one of the other drugs.



Average accuracy per models with Chocolate, Alcohol and Caffeine



Average accuracy per models without Chocolate, Alcohol and Caffeine

Conclusion about the parameters

The histograms show us that to avoid the legal drug is not useful since it decreases the accuracy for all the models except the KNN.

Thus we are keeping our last models and prediction : we take all the variables in consideration, except the one linked to the drug we want to predict (and of course except ID).