Drug Consumption

DATA ANALYSIS AND PREDICTIONS

Louis LIARD / Thomas RIVOAL

Introduction

A dataset about people from all walks of life who answered questions about their drug use has been provided to us.

This is the result of a deep study and therefore it was logic to us to make a meaningful work by responding to a real issue.

Summary of our work:

We wanted to manipulate this fairly complete dataset in the most relevant way to finally obtain the best predictions and especially to be able, in a final perspective, to make prevention to the youngest about the dangers of drug addictions.

Dataset presentation

1885 respondants with 12 known personnality attributes (NEO-FFI-R, BIS-11, ImpSS, level of education, age, gender, country of residence and ethnicity)

They were asked about their last use of 18 legal and illegal drugs

Dataset presentation

The use of drugs is a serious issue and since we want to predict the tendencies of a person to use this or that drug, we have thought carefully beforehand about an optimal preparation of the dataset for better predictions

In this presentation we will explain our thoughts, our doubts, in short everything that pushed us to modify our dataset in our way. There will be three main focuses for the preparation of our dataset :

- ✓ Binary time classification
- ✓ Classification of drugs
- ✓ Cleaning of personality attributes

Binary time Classification

For each drug, the respondants of the study had 7 choices :

"Never Used", "Used over a Decade Ago", "Used in Last Decade", "Used in Last Year", "Used in Last Month", "Used in Last Week", and "Used in Last Day"

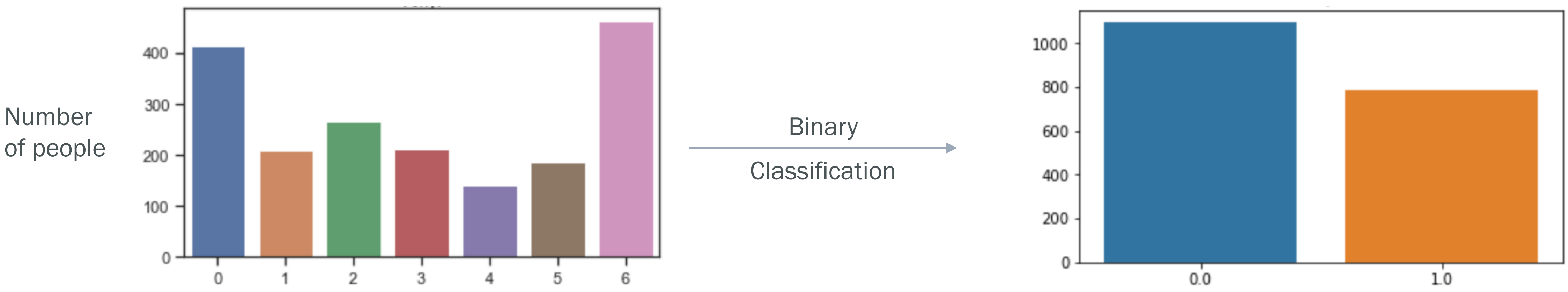
Formally only a participant in the class ‘Never used’ can be called a non-user, but it is not a fondamental definition because a participant who used a drug more than decade ago cannot be considered a drug user for most applications, it might be caused by the curiosity of the youth for example.

The aim here is to define at what point a person is considered a "user" or a "non-user“ that is why we had to create the time scale of our choice

Binary time Classification

After much consideration, we decided to define a "user" as someone who has used a drug in the **last month**

- ✓ Non-users : "Never Used", "Used over a Decade Ago", "Used in Last Decade" and "Used in Last Year"
- ✓ Users : "Used in Last Month", "Used in Last Week", and "Used in Last Day"



Example of this binary classification on cannabis (one of the most balanced)

Drugs are all different in their composition, their way of acting, their relationship to addiction, their power. But we knew that in one way or another we could gather them into groups to simplify our predictions by keeping our relevance.

Thanks to an article, we saw that drug consumption has a ‘modular structure’. As a consequence, we found that drugs could be gathered into 3 groups because the usage of some drugs are significantly correlated. These are the 3 groups whose name is borrowed from their central element :

- ✓ Ecstasy group
- ✓ Heroin group
- ✓ Benzodiazepine group

What is a drug ?

A drug is any psychotropic or psychoactive substance that disrupts the functioning of the central nervous system (sensations, perceptions, moods, feelings, motor skills) or alters states of consciousness.

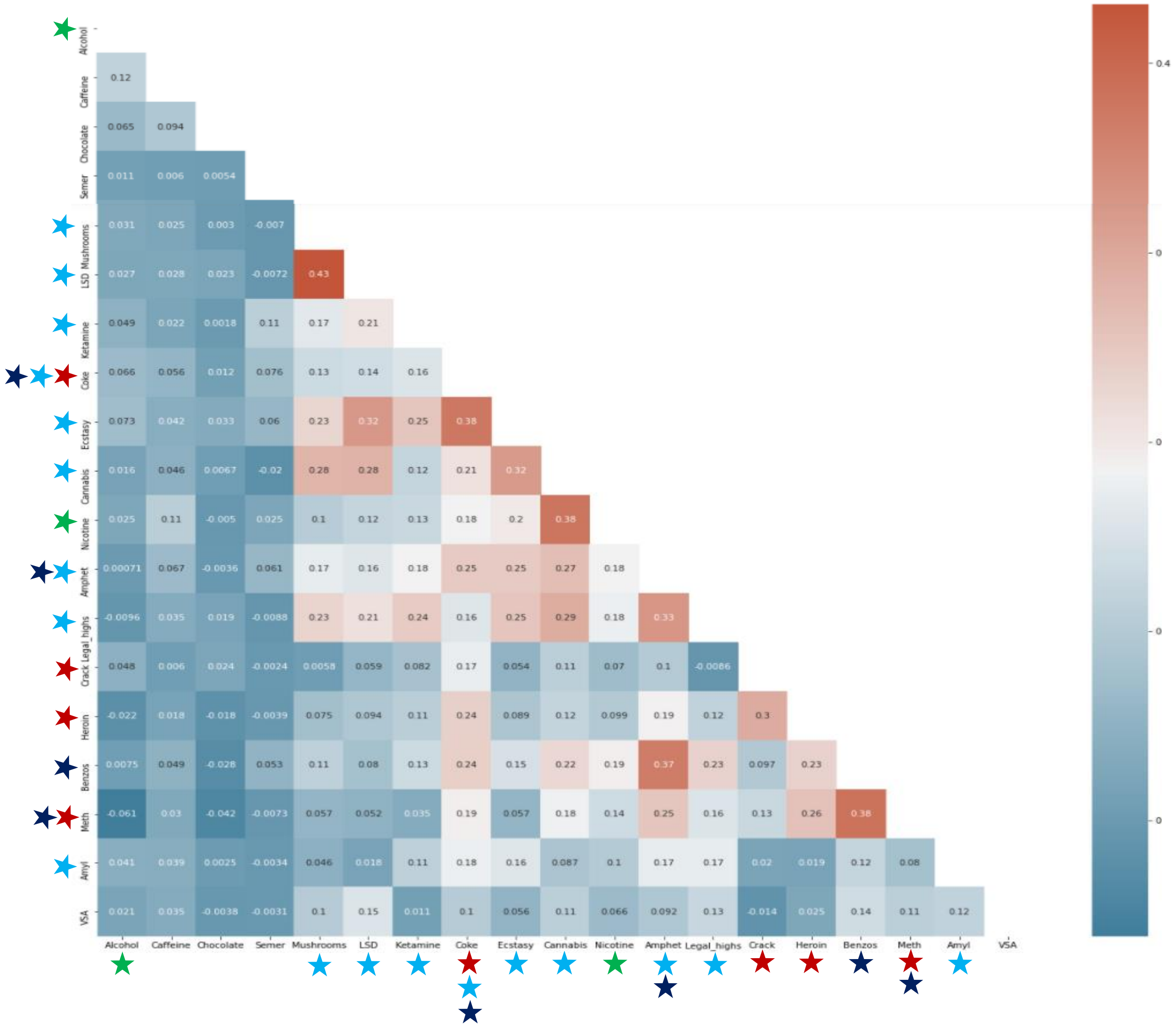
Drug classification

- ✓ Ecstasy group ★
- ✓ Heroin group ★
- ✓ Benzodiazepine group ★

We also wanted to isolate two specific drugs to analyse their use because they are legal, therefore very accessible, but have a very damaging effect on our society :

- ✓ Accessible Goup ★

(Alcohol and nicotine)



	Alcohol	Caffeine	Chocolate	Semer	Mushrooms	LSD	Ketamine	Coke	Ecstasy	Cannabis	Nicotine	Amphet	Legal_highs	Crack	Heroin	Benzos	Meth	Amyl	VSA
Alcohol																			
Caffeine	0.12																		
Chocolate	0.065	0.094																	
Semer	0.011	0.006	0.0054																
Mushrooms	0.031	0.025	0.003	-0.007															
LSD	0.027	0.028	0.023	-0.0072	0.43														
Ketamine	0.049	0.022	0.0018	0.11	0.17	0.21													
Coke	0.066	0.056	0.012	0.076	0.13	0.14	0.16												
Ecstasy	0.073	0.042	0.033	0.06	0.23	0.32	0.25	0.38											
Cannabis	0.016	0.046	0.0067	-0.02	0.28	0.28	0.12	0.21	0.32										
Nicotine	0.025	0.11	-0.005	0.025	0.1	0.12	0.13	0.18	0.2	0.38									
Amphet	0.00071	0.067	-0.0036	0.061	0.17	0.16	0.18	0.25	0.25	0.27	0.18								
Legal_highs	-0.0096	0.035	0.019	-0.0088	0.23	0.21	0.24	0.16	0.25	0.29	0.18	0.33							
Crack	0.048	0.006	0.024	-0.0024	0.0058	0.059	0.082	0.17	0.054	0.11	0.07	0.1	-0.0086						
Heroin	-0.022	0.018	-0.018	-0.0039	0.075	0.094	0.11	0.24	0.089	0.12	0.099	0.19	0.12	0.3					
Benzos	0.0075	0.049	-0.028	0.053	0.11	0.08	0.13	0.24	0.15	0.22	0.19	0.37	0.23	0.097	0.23				
Meth	-0.061	0.03	-0.042	-0.0073	0.057	0.052	0.035	0.19	0.057	0.18	0.14	0.25	0.16	0.13	0.26	0.38			
Amyl	0.041	0.039	0.0025	-0.0034	0.046	0.018	0.11	0.18	0.16	0.087	0.1	0.17	0.17	0.02	0.019	0.12	0.08		
VSA	0.021	0.035	-0.0038	-0.0031	0.1	0.15	0.011	0.1	0.056	0.11	0.066	0.092	0.13	-0.014	0.025	0.14	0.11	0.12	

Drug classification

The article we found suggested three different drug groups with biostatistical arguments.

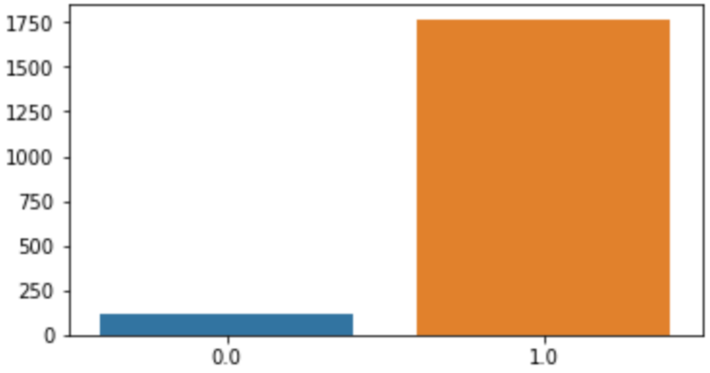
To this, we decided to add another drug using our clustered correlation matrix (previous graph) because we didn't want to leave any drugs behind without thinking.

- ✓ We added amyl nytrite to the heroin group because the drugs with the best correlation are mostly in this group

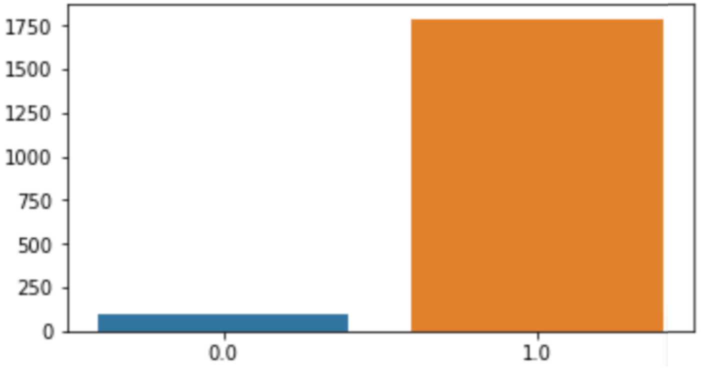
On the other hand, Semer is not useful for predictions, because it is a fake drug created for the study, so we removed it from the dataset

What about the remaining drugs

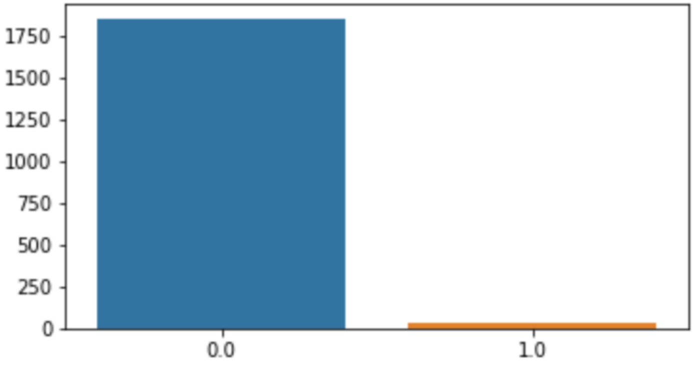
Concerning caffeine, chocolate and VSA, we tried to include them in our accessible group but they showed very unstable results with the binary classification as you can see :



Caffeine



Chocolate



VSA

The accessible group containing only alcohol and nicotine is not very balanced but their presence being very important and damaging at the same time in our society we chose to keep them anyway.

Cleaning of personality attributes

✓ Reminder of the attributes of our dataset

7 personality traits :

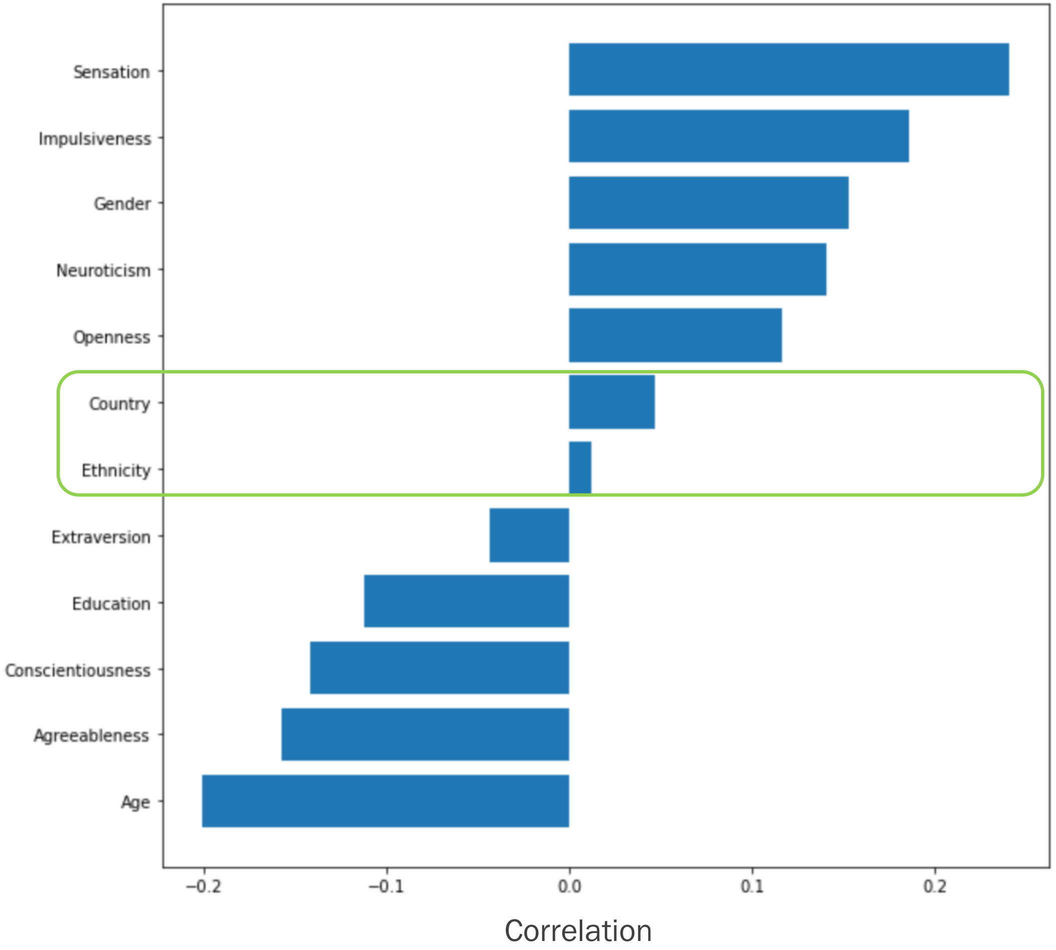
NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, conscientiousness), BIS-11 (impulsivity), ImpSS (sensation seeking)

5 General characteristics :

level of education, age, gender, country of residence and ethnicity.

In our opinion and even scientifically, the first seven personality traits are necessarily to be taken into account. They are very correlated to drug consumption and as a consequence we did not remove or modify any of them.

After having studied the role and importance of attributes in the dataset, we found that for the three groups were very poorly correlated to ethnicity and the country.



Here is an example of a correlation graph for the heroin group with the different attributes :

We obviously understand that the country and the ethnicity are the less correlated/relevant features.

As the graphs are roughly the same for the 3 groups we removed those attributes.

Cleaning of personality attributes

Brief summary

Current status of our database

✓ Monthly-base binary classification (User / Non-user)

✓ Relevant clustering of drugs into 3 main groups based on a correlation of drug use :

- Heroin pleiad : crack, cocaine, methadone, and heroin
- Ecstasy pleiad : amphetamines, cannabis, cocaine, ketamine, LSD, magic mushrooms, legal highs, ecstasy and amyl
- Benzodiazepines pleiad : methadone, amphetamines, benzodiazepine and cocaine

✓cleaning of personality attributes : removal of ethnicity and country of residence

Predictions Summary

Thus, through our prediction research work, we were able to build three Machine Learning models to predict whether an individual will potentially become a drug user.

We found that the LogisticRegression model was the best performing model for predicting Ecstasy and Benzo drug use.

On the other hand, the RandomForestClassifier performed best in predicting Heroin-type drugs.

At the end, we obtained 3 ML models that performed well thanks to a GridSearch.

Conclusion

With this dataset well set up for predictions, we invite you to go on our Django page to see if you are likely to use this or that drug.

For more information on our coding methods, please check our jupyter notebook, all the resources are available on our github link.