STUDY PROJECT REPORT : Drug Consumption

By
Chloé Coursimault
Héloïse de Castelnau

DataSet Exploration

Our dataset from the **UCI Machine Learning** repository is the result of an online survey conducted between 2011 and 2012 among **1,885 respondents** aged 18 and over, from English-speaking countries.

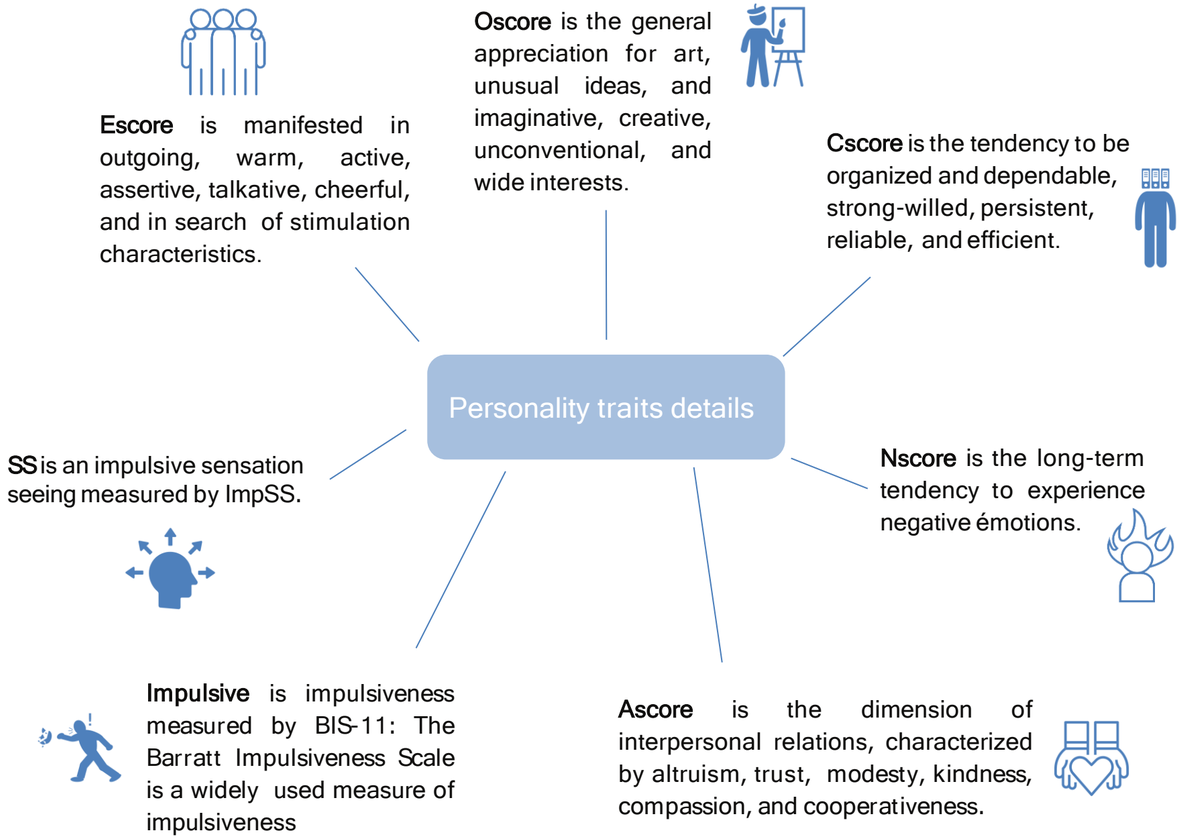
The survey collected data including Big Five personality traits (NEO-FFI-R), impulsivity (BIS-11), sensation seeking (ImpSS), and demographic information.

The data set contained information on the consumption **of 18 central nervous system psychoactive drugs legal and illegal**. For each of these drugs, each individual had to choose his level of consumption :

- CL0 Never Used
- CL1 Used over a Decade Ago
- CL2 Used in Last Decade
- CL3 Used in Last Year
- CL4 Used in Last Month
- CL5 Used in Last Week
- CL6 Used in Last Day

Demographic features	Personality traits	Legal Drugs	Illegal Drugs
• Age	• Nscore	• Alcohol	• Amphet
• Gender	• Escore	• Caff	• Amyl
• Education	• Oscore	• Chocolate	• Benzos
• Country	• Ascore	• Nicotine	• Cannabis
• Ethnicity	• Cscore		• Crack
	• Impulsive		• Ecstasy
	• SS		• Heroin
			• Ketamine
			• Legalh
			• LSD
			• Meth
			• Mushrooms
			• VSA

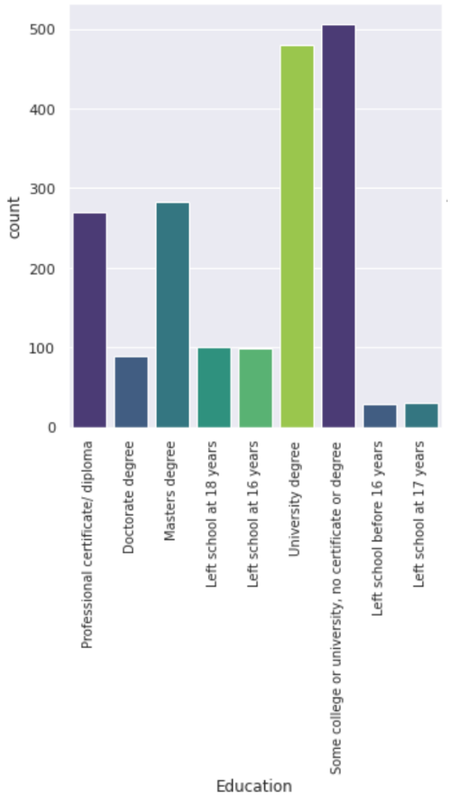
Furthermore, to detect Fraud, the survey organizers added a fake drug "sumer".

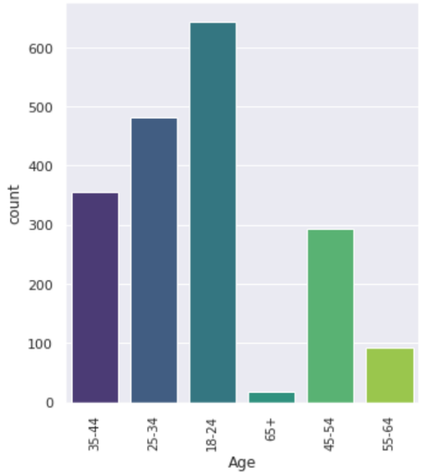


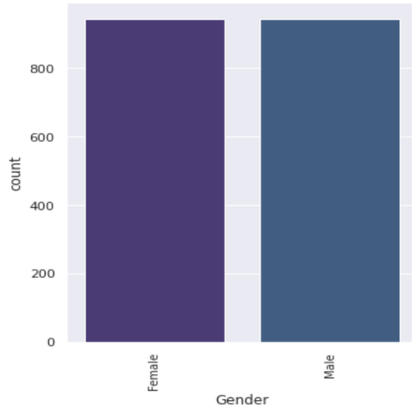
DataSet Analysis

DEMOGRAPHIC DATA

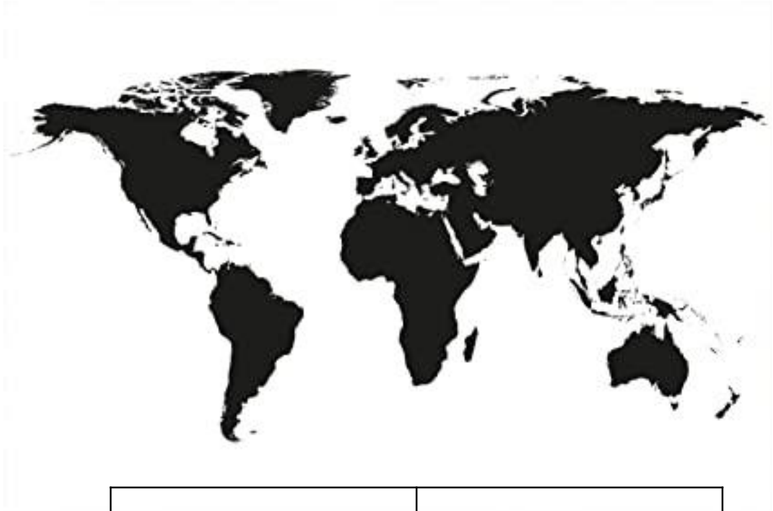
Some categories are almost uniformaly disparate such as **Education** and **Age** or even equally represented wich **Gender**.



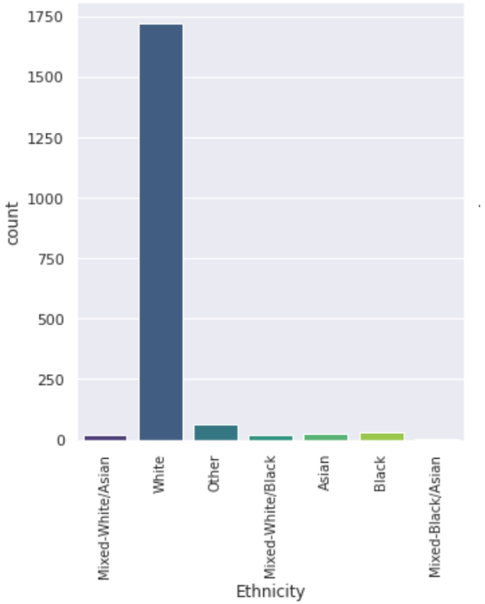




We can observe some unevenly represented categories within the columns of the Dataframe, suchs as **Ethnicity** that count up to 90% of white or **Country** with mostly UK participants. This biais may false our predictions so we can already say that we will get rid of thoses ubalanced goups distribution.

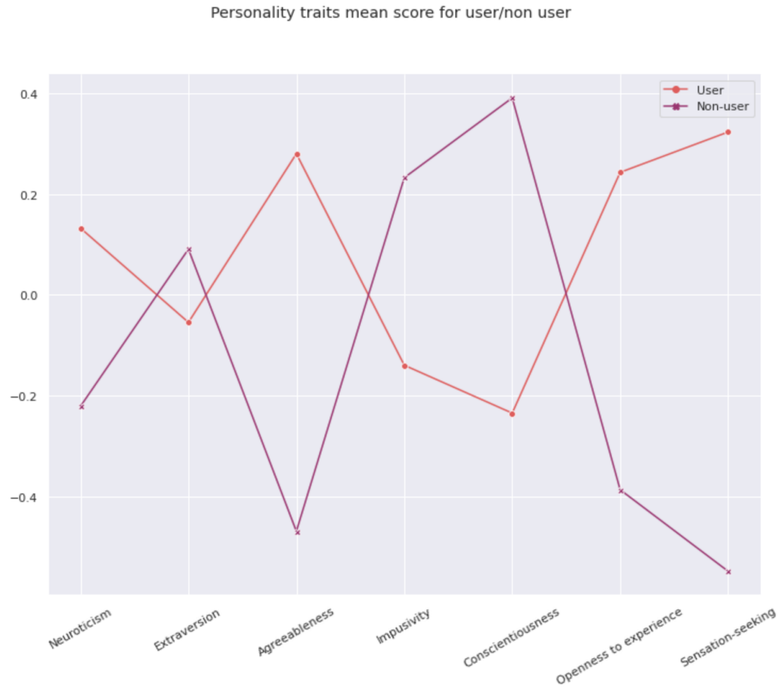


Country	%
UK	55.38
USA	29.55
Other	6.26
Canada	4.61
Australia	2.86
Republic of Ireland	1.06
New Zealand	0.26



INTERESTINGS OBSERVATIONS

By creating a new column of **user/non-user** based on the annual consumption of illegal drugs, we ploted the mean score for each **personality traits** for cosumer or not consumers.We can see a strong correlation between the consumption and the changing behaviour, the personality traits are **diametricelly opposed**. It's interesting because we may argue that it's drugs use that may change the behaviour of the participants. Because the question of how personality change would affect how we will develop our ML model later.

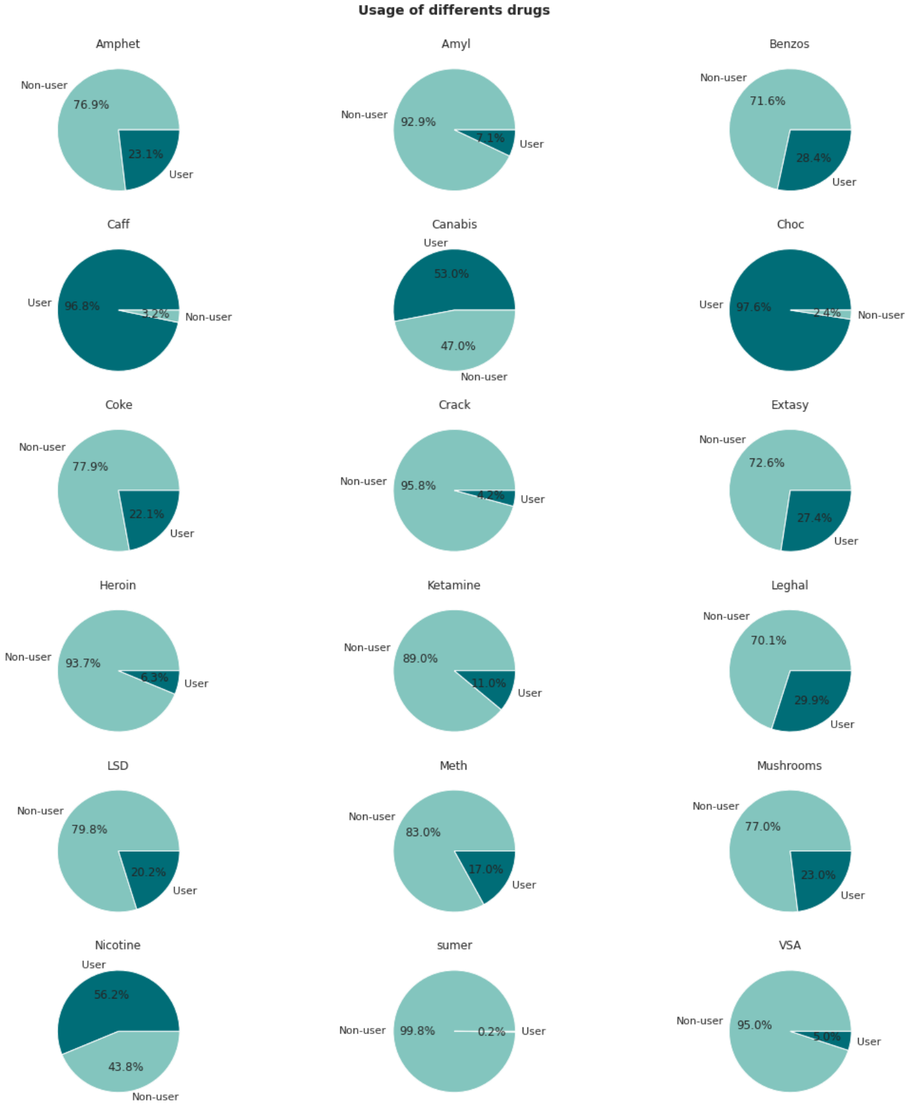


We can observe a widely disparate representation of user/non-user for illegal and legal drugs.

Fisrt, we can spot a small proportion of cheater 0.2%, but because it may add noise we will get rid of this column in our futur prediction.

Then we can see that the illegal drug the most used is the Cannabis with up to 47% of users, this drug is weel-known as a Getaway drug.

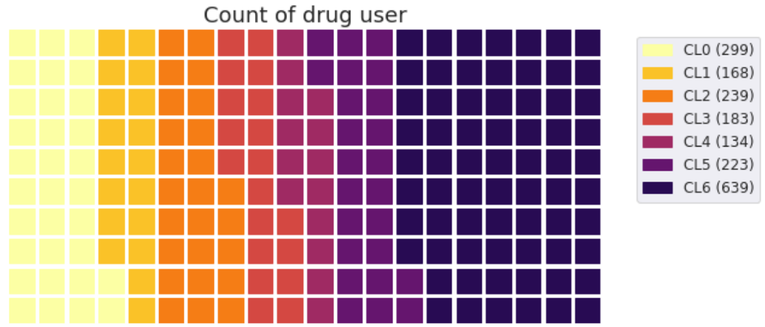
A gateway drug is a habit-forming drug that can lead to the use of other, more addictive drugs. They include alcohol, marijuana, nicotine, and prescription drugs.

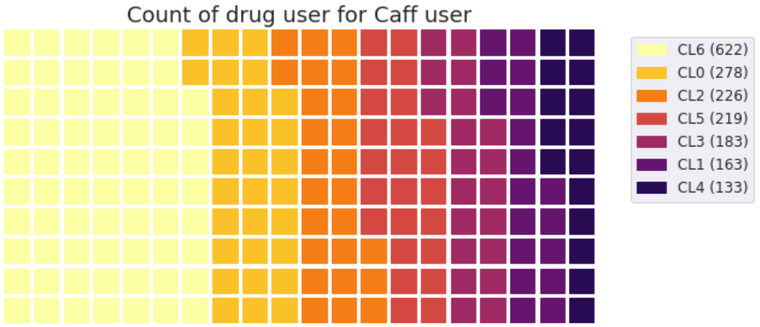


INTERESTINGS OBSERVATIONS (2)

Normal use

We ploted the proportion of different categories of drugs usage within the dataset. We can see a widely disparate representation of thoses classes, with a majority of users up to a year (rf Machine Learning part). Before studying the Getaway drugs we also poted a waffle chart for Caffeine as a referent.

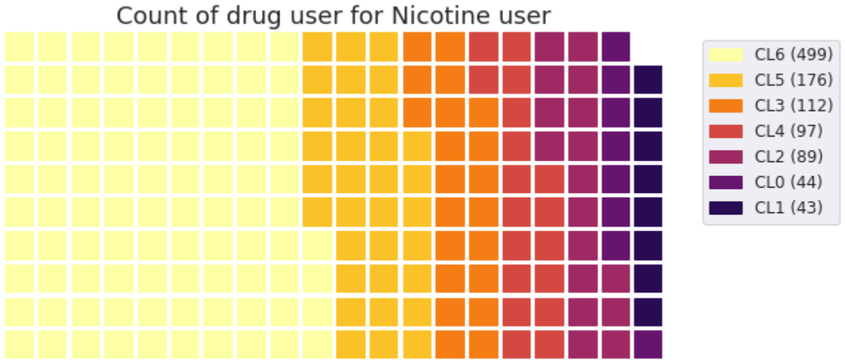


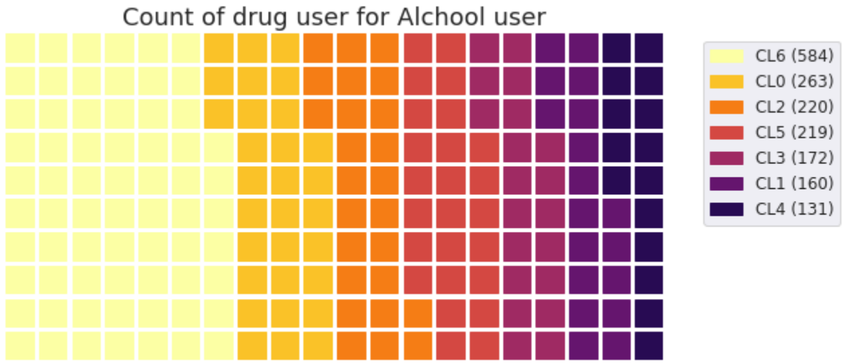


Getaway drugs

We can observe here that the *Getaway drugs* **Nicotine** & **Alchool** seems to have a more highly count of illegal drug user.
However, we didn't get the timeline for each participants of their consumption, so it's complicated to say if the use of alchool could impact the use of Amphets for example.

We will see thoses correlations more detailled in the Machine Learning part, but it do seems that the consumption of Nicotine or Alchool is *strongly correlated* with the consumption of others illegal drugs.





Machine Learning

Choosen Output

MultiClass prediction:

How much the individual with the characteristics X used illegal drugs lately (up to one year) ?

We created a category « Illegal drug use » wich take for each participant withing the illegal drugs column his highest consumption. We tried to make predictions on this output using Logistic Regression and Neural network but got mixed results. We thought that our dataset is unbalanced (it has more samples for some classes than others). This can make the classifier biased toward the one or two classes with lost of samples. We also made the hypothesis that our dataset was not specific enough.

Binary Classification

Has the individual with the characteristics X used illegal drugs lately (up to one year)?

CL0
CL1
CL2
→ Non-user

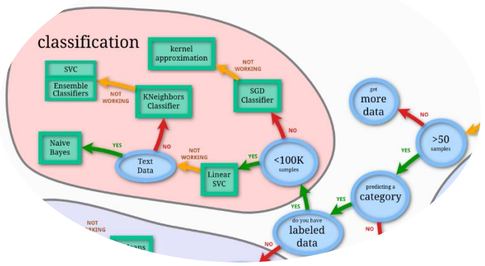
CL3 Used in Last Year
CL4 Used in Last Month
CL5 Used in Last Week
CL6 Used in Last Day
→ User

	ID	Age	Gender	Education	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS	Alchool	Caff	Choc	Nicotine	Illegal drug use	Illegal drug user/non-user
0	1	0.49788	0.48246	-0.05921	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1.18084	0	0	0	0	CL2	Non-user
1	2	-0.07854	-0.48246	1.98437	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575	0	0	0	0	CL4	User
2	3	0.49788	-0.48246	-0.05921	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0.40148	0	0	0	0	CL3	User
3	4	-0.95197	0.48246	1.16365	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084	0	0	0	0	CL3	User
4	5	0.49788	0.48246	1.98437	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	-0.21575	0	0	0	0	CL3	User
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

What Algotihms did we implemented?

We tried multiple Machine Learning Algorithms from the scikit-learn library with the « Illegal drug user/non-user » as an output, by checking for each individual if he used an illegal drug up to a year before the survey.
Here is a little overview of the algorithms we tested :

- Logistic Regression
- Linear SVC
- Gaussian NB
- Random Forest
- Decision Tree
- KNN
- XGBOOST
- SVM
- Neural Network



Here is a little cheat sheet of the scikit-learn library.

Algorithms predictions

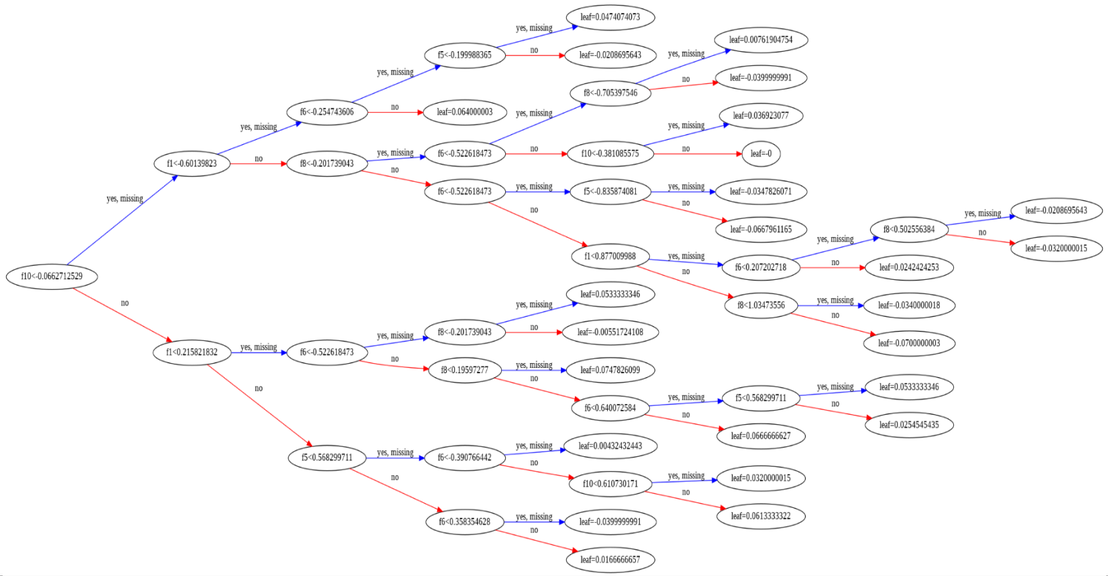
How does it work ? - a short overview

✓ XGBOOST, Linear Regression & Neural network

XGBOOST-

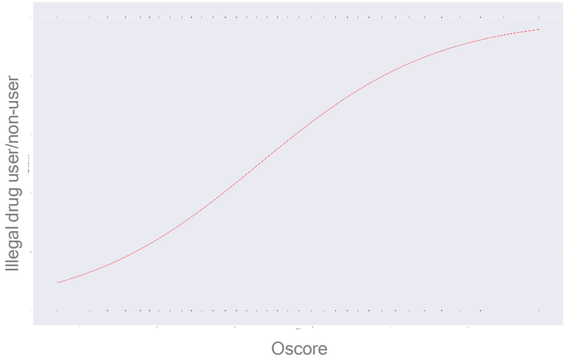
We employ XGBoost, a fast and strong gradient boosting framework that is also based on multiple decision tree learning, in this part. Gradient boosting is a technique for transforming weak learners into strong ones. When combined with prior models, it is based on the premise that predicting the next tree inside a model can minimize prediction errors.
XGBoost, like Random Forest, provides a number of benefits, including: trains quicker and more effectively while consuming less memory.In general, very efficient, especially on huge datasets, with more complicated trees using a level-wise level-wise growth strategy.

You can see that variables are automatically named like f1 and f5 corresponding with the feature indices in the input array.
You can see the split decisions within each node and the different colors for left and right splits (blue and red).



Logistic regression Regression-

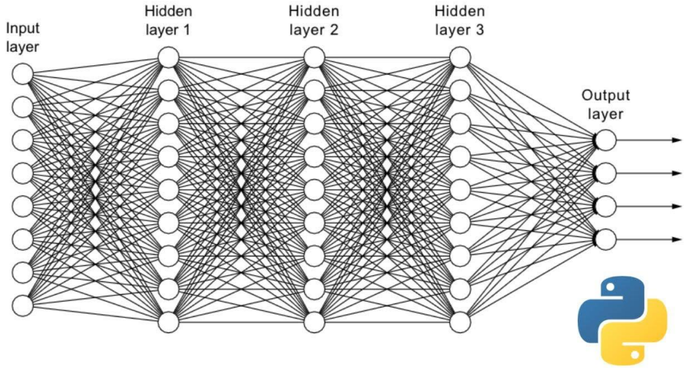
For the base-line model, we consider the multiple linear regression model from sklearn.linear_model library LinearRegression.
statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).



Neural Network-

In essence, neural networks are self-optimizing functions that map inputs to the desired outputs. The function may then take a fresh input and predict an output based on the function it developed with the training data.
An ANN is made up of the following elements:

- An input layer that receives data and pass it on
- A hidden layer
- An output layer
- Weights between the layers
- A deliberate activation function for every hidden layer.



✓ Remark -
For each of thoses algorithms we used a GridSearch with cross-validation to select the hyperparameters that are most suited. It entails putting a machine learning model through many permutations of hyperparameters in order to assess its performance and make improvements.

Conclusion

Which ones are the best model ?

What metrics did we used ?

The goal of our predictions is to predict if someone is going to be a drug user or not, so we want to reduce the False positive rates. Therefore we want to maximize our recall in attempts to answer to the following question What proportion of actual positives was identified correctly?

$$\text{Recall} = \frac{TP}{TP + FN}$$

We will also keep an eye on the other metrics such as the accuracy, the precision and in extension the f1-score (wich is the report between recall and precision)