Drug consumption analysis

Python for Data Analysis

ESILV A4 - DIA 2
Kévin CELIE - Mariyam CHEICK ISMAIL
Referent teacher : Mr. Benjamin BEJBAUM

Table of contents

Content of the repository

- Two notebooks
  - DRUG_Analysis for the analysis
  - DRUG_Prediction for the modelisation
- Two python scripts
  - app.py for the flask server
  - request.py for the requests we do
- Charts folder where charts are exported in .png
- Models folder to store all models

Dataset

- Describe
- Clean
- Explore

Describe the dataset

Database contains records for 1885 respondents.

For each respondent 12 attributes are known:

- Personality measurements which include NEO-FFI-R
  - Neuroticism,
  - Extraversion,
  - Openness to experience,
  - Agreeableness, and
  - Conscientiousness),
- BIS-11 (impulsivity),
- and ImpSS (sensation seeking),
- level of education,
- age,
- gender,
- country of residence and
- ethnicity.

Describe the dataset

Drugs :

- Alcohol,
- Amphetamines,
- Amyl nitrite,
- Benzodiazepine,
- Cannabis,
- Chocolate,
- Cocaine,
- Caffeine,
- Crack,
- Ecstasy,
- Heroin,
- Ketamine,
- Legal highs,
- LSD,
- Methadone,
- Mushrooms,
- Nicotine,
- Volatile Substance Abuse (VSA) and
- Semeron, one fictitious drug which was introduced to identify over-claimers.

Describe the dataset

For each drug they have to select one of the answers:

- "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".

Database contains 18 classification problems. Each of independent label variables contains one of each of these seven classes.

Cleaning the dataset

The dataset presents multiple values, which makes it hard to read and analyse. In order to understand values we have, we need to clean the dataset.

But before cleaning it, we must check if there is any missing values. Hopefully, we don't have any here.

Then, we can proceed the cleaning of the dataset.

	ID	AGE	GENDER	EDUCATION_LEVEL	COUNTRY	ETHNICITY	NSCORE_VALUE	ESCORE_VALUE	OSCORE_VALUE	ASCORE_VALUE	...	ECSTASY_CONSUMPTION
0	1	0.49788	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	...	CL0
1	2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	...	CL4
2	3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	...	CL0
3	4	-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	...	CL0
4	5	0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	...	CL1
5	6	2.59171	0.48246	-1.22751	0.24923	-0.31685	-0.67825	-0.30033	-1.55521	2.03972	...	CL0
6	7	1.09449	-0.48246	1.16365	-0.57009	-0.31685	-0.46725	-1.09207	-0.45174	-0.30172	...	CL0
7	8	0.49788	-0.48246	-1.73790	0.96082	-0.31685	-1.32828	1.93886	-0.84732	-0.30172	...	CL0
8	9	0.49788	0.48246	-0.05921	0.24923	-0.31685	0.62967	2.57309	-0.97631	0.76096	...	CL0
9	10	1.82213	-0.48246	1.16365	0.96082	-0.31685	-0.24649	0.00332	-1.42424	0.59042	...	CL0

10 rows × 32 columns

Before cleaning

	ID	AGE	GENDER	EDUCATION_LEVEL	COUNTRY	ETHNICITY	NSCORE_VALUE	ESCORE_VALUE	OSCORE_VALUE	ASCORE_VALUE	...
0	1	35-44	F	Professional certificate/diploma	UK	Mixed-White/Asian	39	35	40	29	...
1	2	25-34	M	Doctorate degree	UK	White	29	51	53	40	...
2	3	35-44	M	Professional certificate/diploma	UK	White	31	44	38	24	...
3	4	18-24	F	Masters degree	UK	White	34	33	44	39	...
4	5	35-44	F	Doctorate degree	UK	White	43	27	41	33	...
5	6	65+	F	Left school at 18 years	Canada	White	29	37	33	47	...
6	7	45-54	M	Masters degree	USA	White	31	31	41	33	...
7	8	35-44	M	Left school at 16 years	UK	White	24	51	38	33	...
8	9	35-44	F	Professional certificate/diploma	Canada	White	42	54	37	40	...
9	10	55-64	M	Masters degree	UK	White	33	39	34	39	...

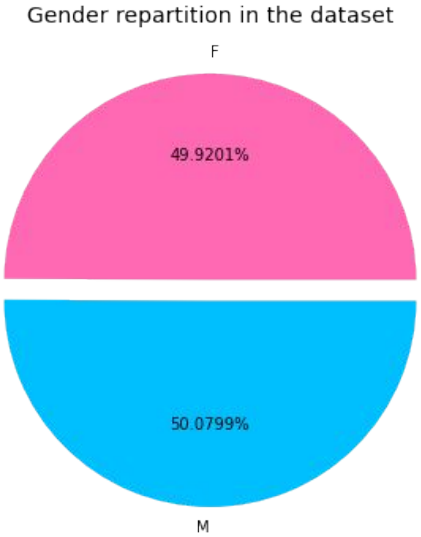
10 rows × 32 columns

After cleaning it, makes it more understandable

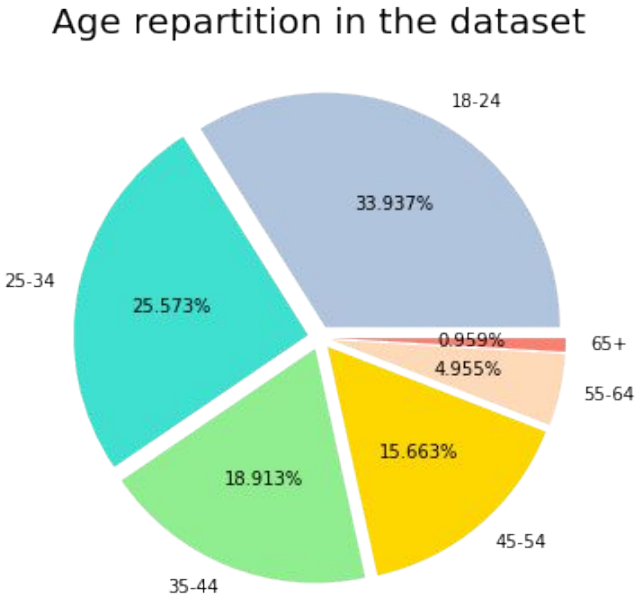
Exploring the dataset

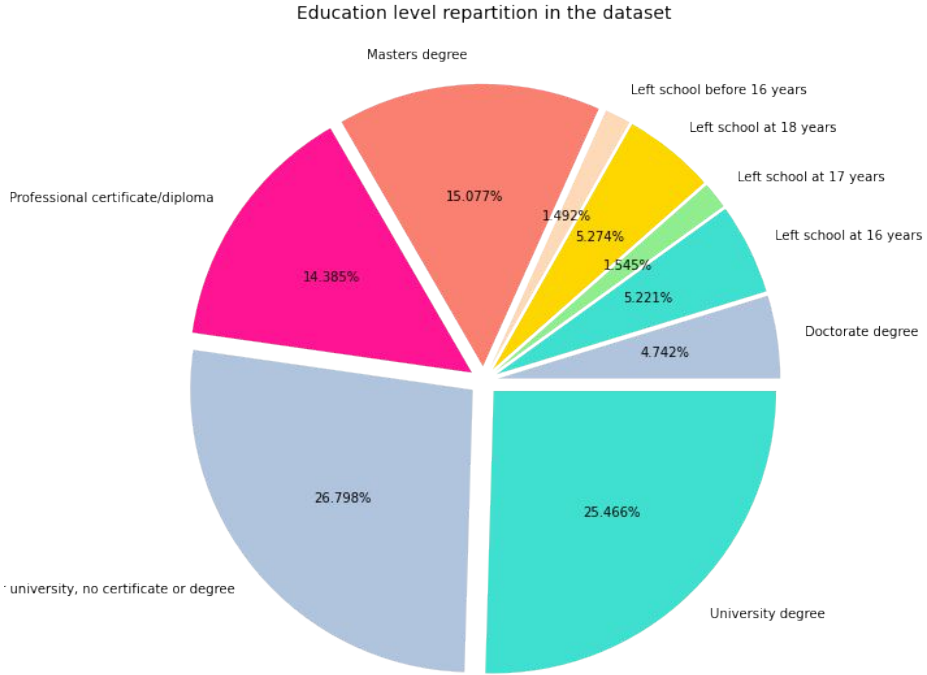
This dataset gives information about people : their gender, their age, their education level, their ethnicity and their living country.

In this dataset, we have almost a perfect parity :
with 50,08% of men
and 49,92% of women.



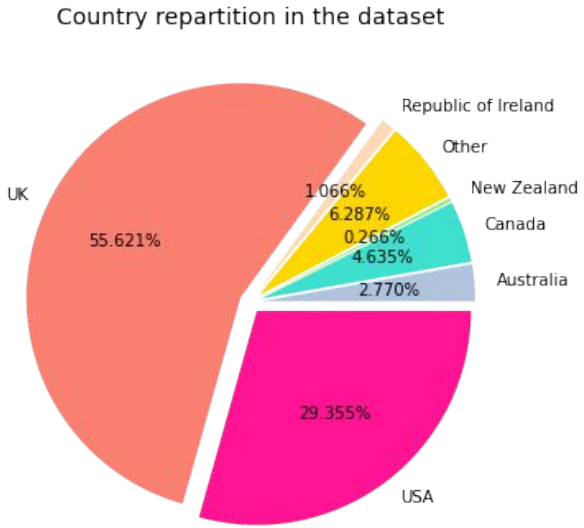
Also, most of the people went to the university.

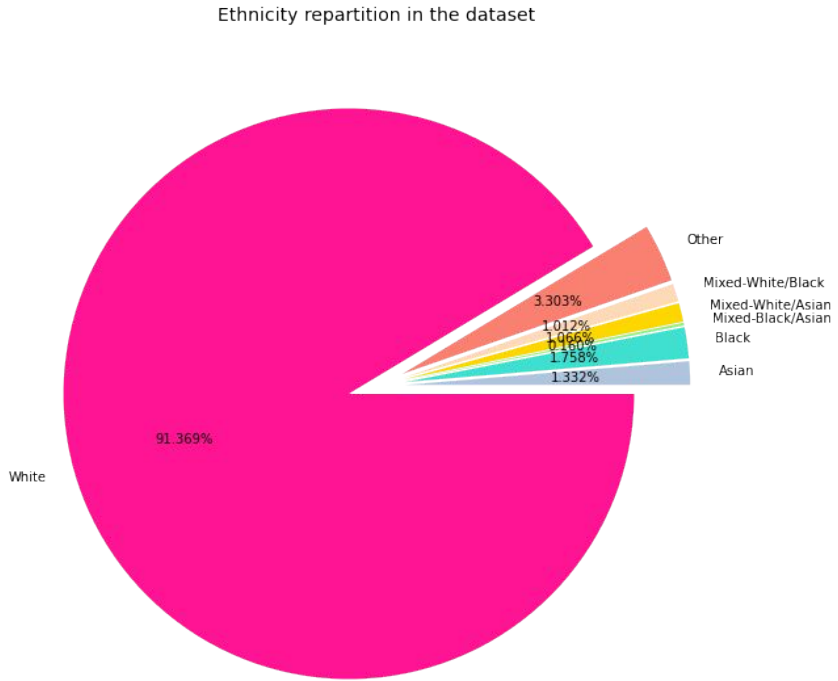




This dataset is mostly composed of people aged between 18 and 44 years old.

There are almost only white people, which is not really representative but expectable regarding to the country living repartition





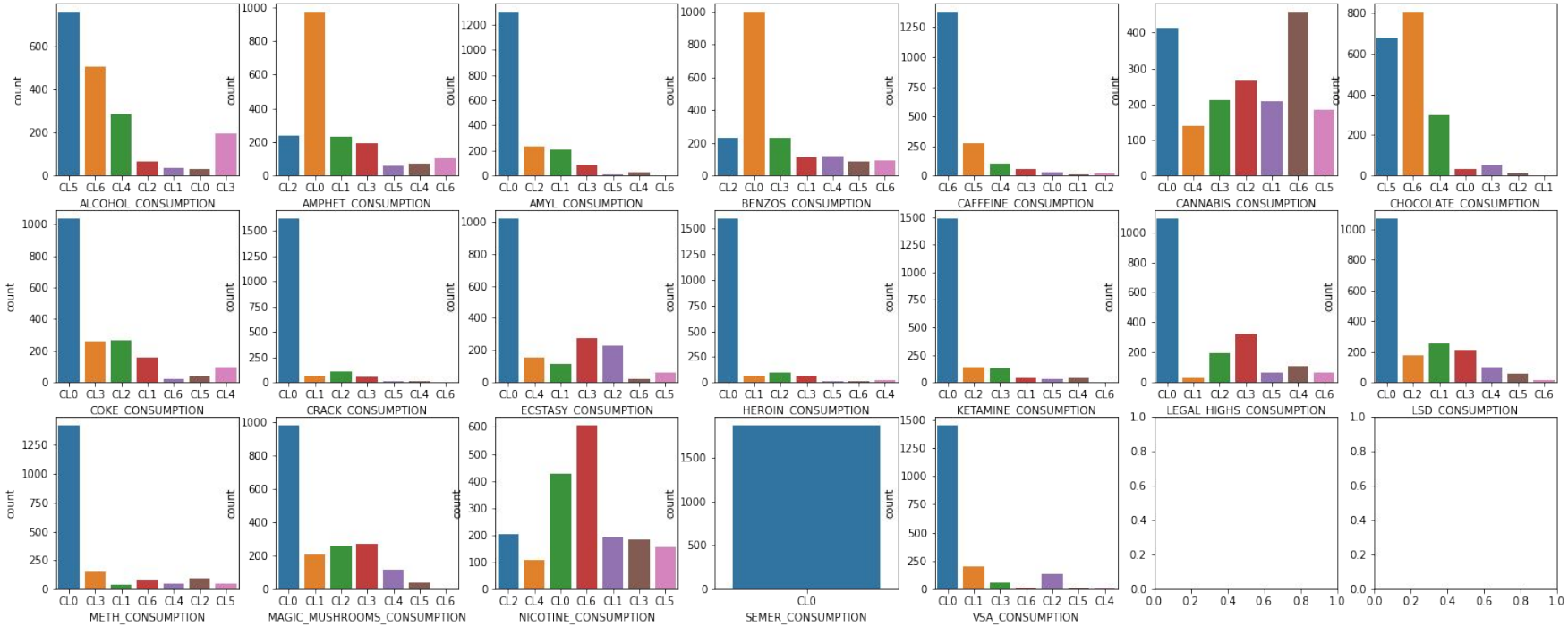
This dataset is mainly composed of people from the UK and the USA.

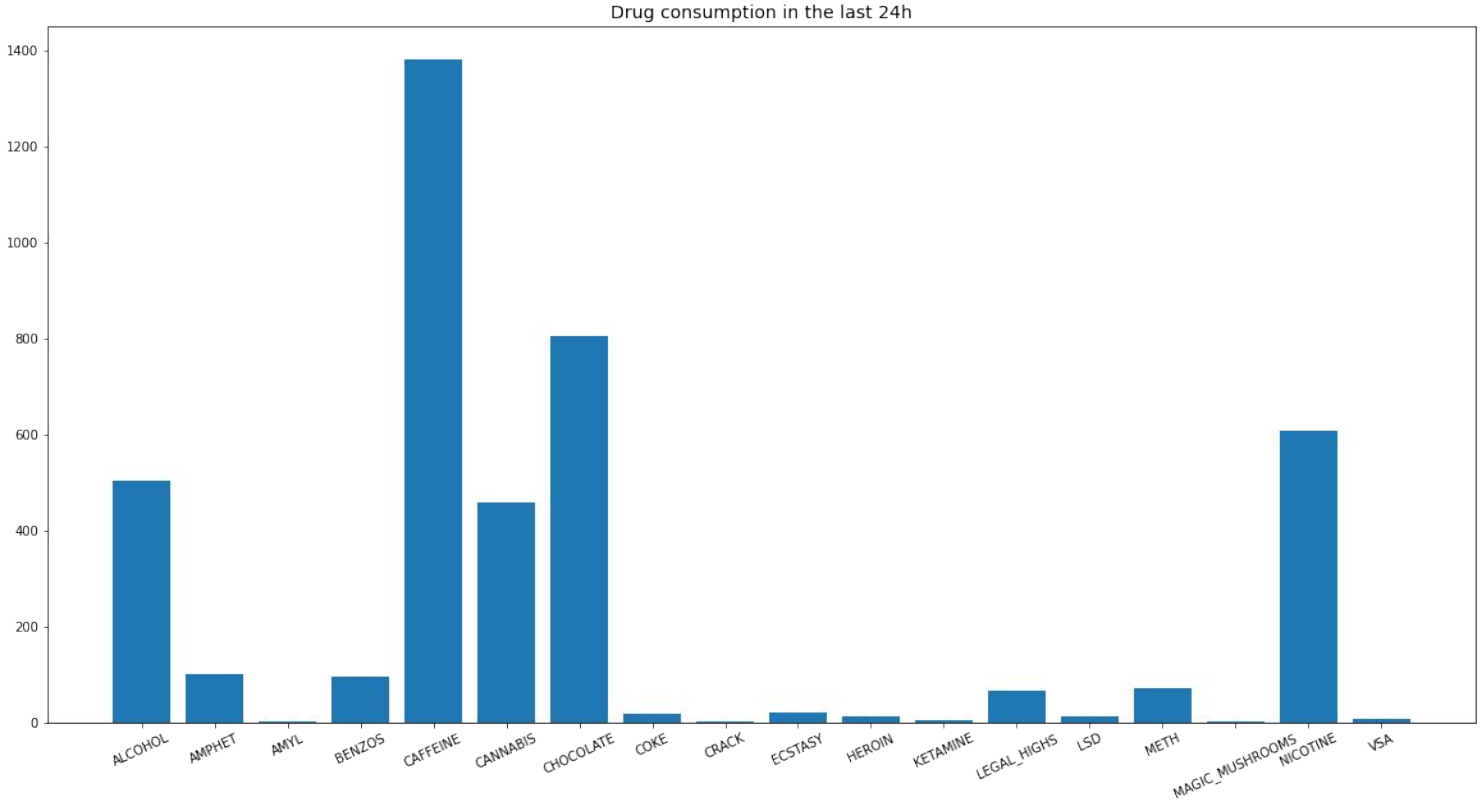
Not worldwide representative

Analysis : the drug consumption

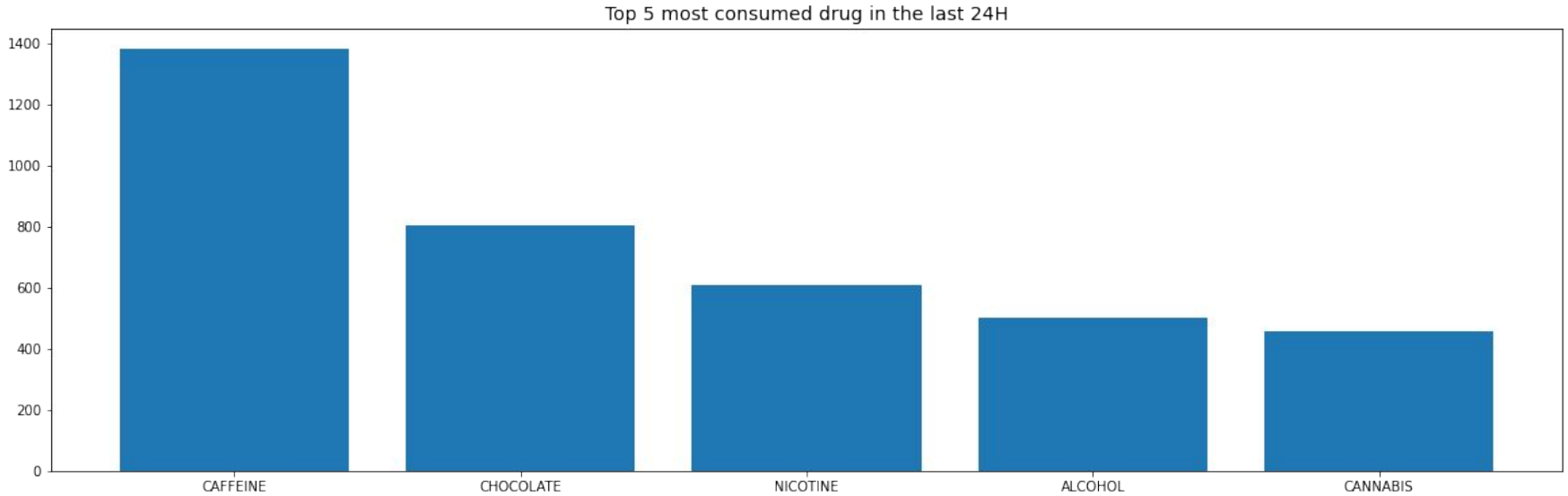
- In general
- In the last 24 hours
- Top 5 and Top 10
- The drug addict 💉
- The little angel 😇
- The liars 😦

Drug consumption in general

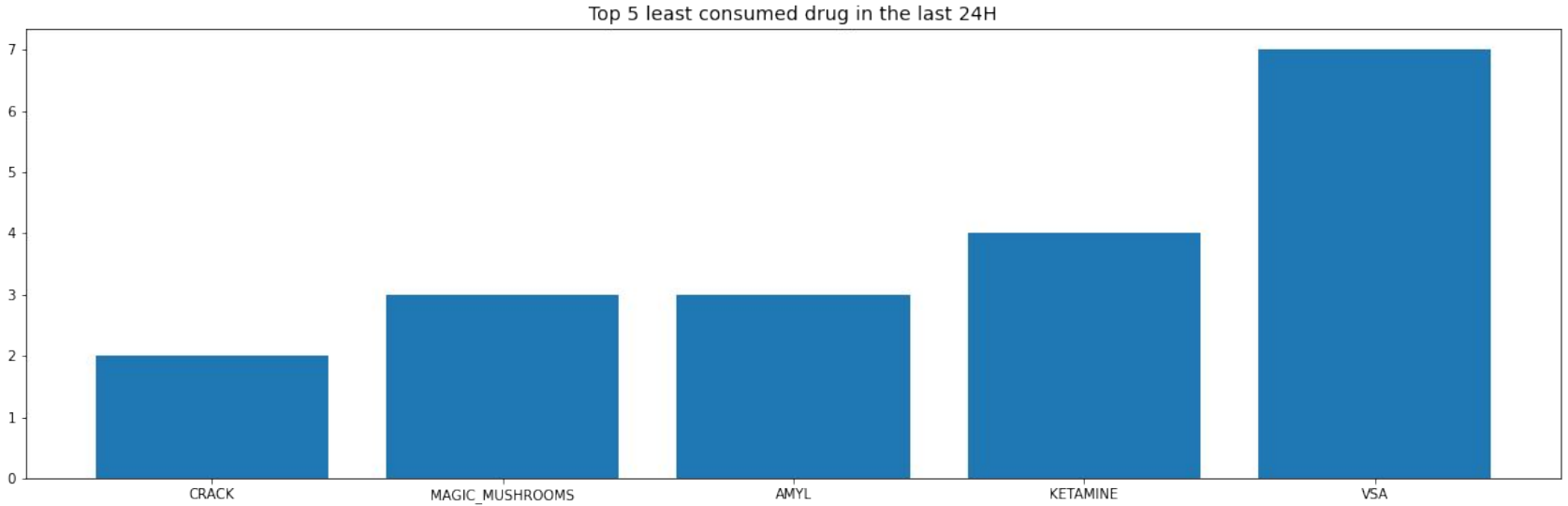




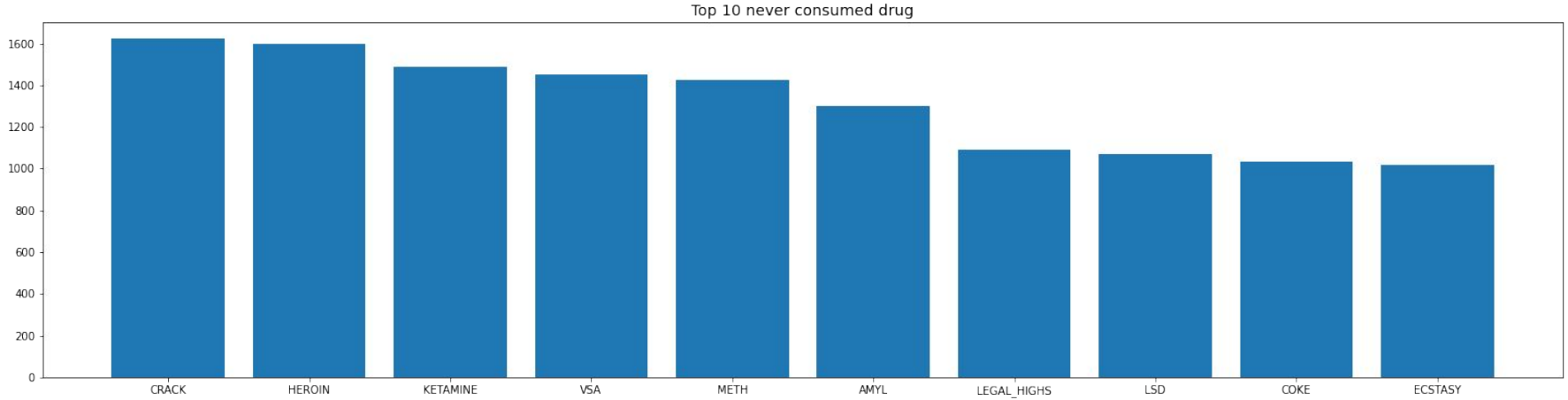
Drug consumption in the last 24h



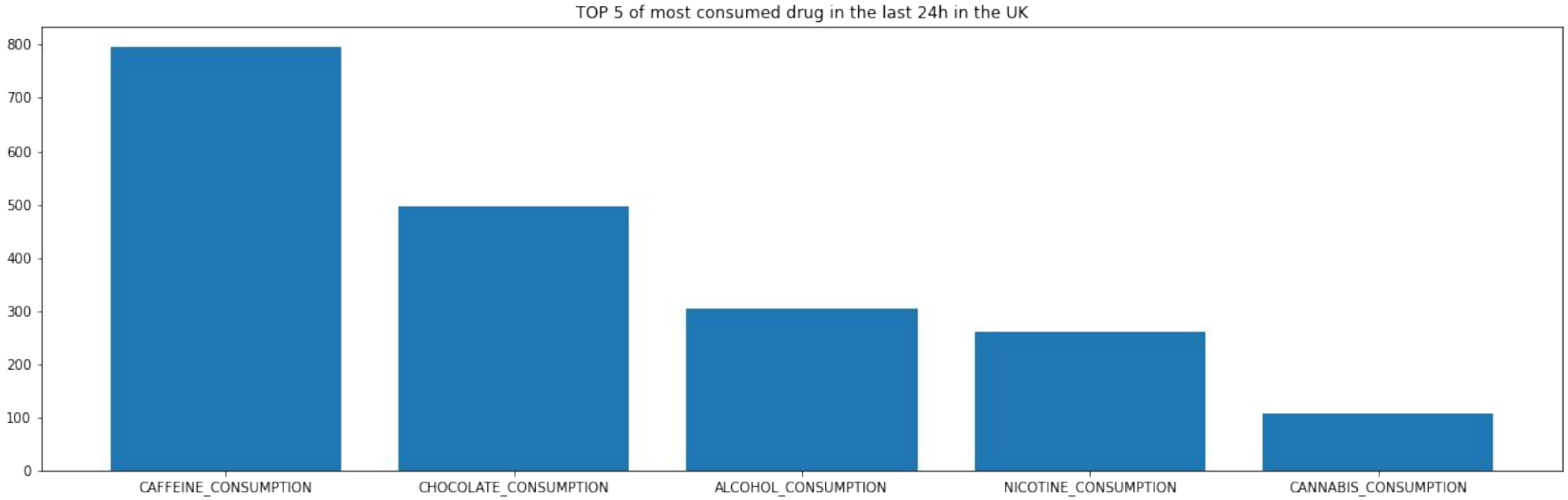
Top 5 - Most consumed drug in the last 24h



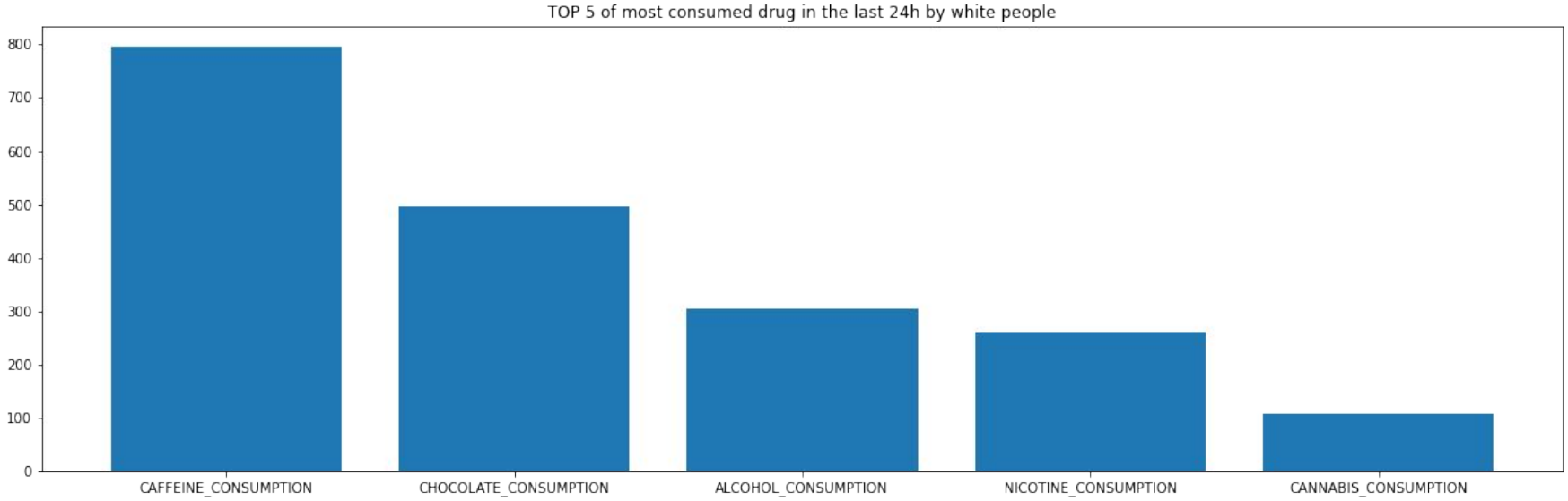
Top 5 - Least consumed drug in the 24h



Top 10 - Never consumed drug



Top 5 - Most consumed drug in the UK in the last 24h



Top 5 - Most consumed drug by white people in the last 24h

The drug addict - who has consumed the most drug in the last 24h : 8 drugs

ID	AGE	GENDER	EDUCATION LEVEL	COUNTRY	ETHNICITY
983	18-24	M	Some college or university, no certificate or degree	UK	White

The little angel - who has never consumed any drug at all*

ID	AGE	GENDER	EDUCATION LEVEL	COUNTRY	ETHNICITY
1281	18-24	M	University degree	UK	Asian

*I think he lies, or his life is meaningless... He has never consumed chocolate or caffeine in his life ? That's too sad, he really misses something

😵 The liars : they have been removed for the analysis

They claimed to have consumed a fictitious drug, the Semeron.

ID	AGE	GENDER	EDUCATION LEVEL	COUNTRY	ETHNICITY
730	25-34	F	Left school at 16 years	Australia	White
821	18-24	M	Some college or university, no certificate or degree	Australia	Asian
1520	18-24	M	Some college or university, no certificate or degree	USA	White
1537	18-24	F	Some college or university, no certificate or degree	USA	Other
1702	35-44	F	University degree	USA	White
1773	18-24	M	Left school at 18 years	USA	Mixed-White/ Black
1810	18-24	F	Left school at 17 years	USA	White
1827	18-24	F	University degree	USA	White

Modeling for predictions

- Problem
- Models
- Our choice
- Export model
- BONUS

Problem

Problem which can be solved:

- Seven class classifications for each drug separately.
- Problem can be transformed to binary classification by union of part of classes into one new class. For example, "Never Used", "Used over a Decade Ago" form class "Non-user" and all other classes form class "User".
- The best binarization of classes for each attribute.
- Evaluation of risk to be drug consumer for each drug.

We have decided to choose the first problem : the **seven-class classification for each drug.**

Models

- Support Vector Machine for Classification : the Support Vector Clustering (SVC)

In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis.
SVMs are one of the most robust prediction methods.

- Decision Trees and Random Forests

Decision tree builds classification or regression models in the form of a tree structure.
It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
The final result is a tree with decision nodes and leaf nodes.
Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier.
Random forest adds additional randomness to the model, while growing the trees.

Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features.

Models

- K-Nearest Neighbours (KNN)

An approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in.

- Deep Learning with tensorflow : Neural Network

Neural networks are a set of algorithms, that are designed to recognize patterns.
They interpret sensory data through a kind of machine perception, labeling or clustering raw input.

Our choice

Even if the SVC model was the most robust one,
we have decided to take **Random Forest** for our modeling part.

Our Random Forest has these parameters :

- ➔ bootstrap='True',
- ➔ max_depth=15,
- ➔ min_samples_leaf=5,
- ➔ n_estimators=10,
- ➔ random_state=1

Export models

To export our models, we have used “pickle”, a module implements binary protocols for serializing and de-serializing a Python object structure.

> *“In computing, **serialization** (US spelling) or **serialisation** (UK spelling) is the process of translating a data structure or object state into a format that can be stored (for example, in a file or memory data buffer) or transmitted (for example, across a computer network) and reconstructed later (possibly in a different computer environment).*
>
> *When the resulting series of bits is reread according to the serialization format, it can be used to create a semantically identical clone of the original object.”*

Source : Serialization

Bonus : Neural Network

- Accuracy stuck at 50% during training.

```
48/48 [==============================] - 0s 2ms/step - loss: 0.0220 - accuracy: 0.5073
```

- An unconvincing model.

```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] expected [1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0]
[1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] expected [1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0]
[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0] expected [0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 0]
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] expected [0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0]
[1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] expected [1 1 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0]
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0] expected [0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0]
```

Web API

- Presentation of the framework
- What is an API ?
- Our API

The framework : Flask

Flask is a micro web framework written in Python.

It was created by Armin Ronacher, the 1st of April, 2010. It classified as a microframework because it is very light and do not require any other library.

What is an API ?

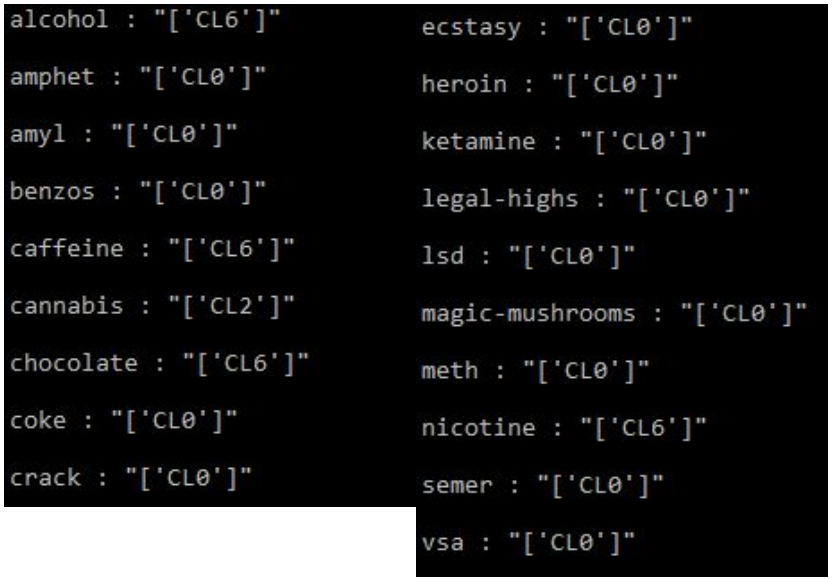
An **application programming interface** ('API') is a computing interface that defines interactions between multiple software intermediaries. It defines the kinds of calls or requests that can be made, how to make them, the data formats that should be used, the conventions to follow, etc. It can also provide extension mechanisms so that users can extend existing functionality in various ways and to varying degrees.

An API can be entirely custom, specific to a component, or designed based on an industry-standard to ensure interoperability. Through information hiding, APIs enable modular programming, allowing users to use the interface independently of the implementation.

Our API

Passing a JSON via a POST request in the following format :

{“person” : [*Age, Gender, Education, Country, Ethnicity, Nscore, Escore, Oscore, Ascore, Cscore, Impulsive, SS*]}



Conclusion

Conclusion : Possible improvements

- For the modeling part :
  - The accuracy scores are not really great and for some are even too bad, even with grid search.
    - Possible solutions : Try other models and change the split ratio for training and testing set

- For the API part :
  - We cannot test the API using the browser.

Where to find this project

This project is available on Github :
https://github.com/m-cheicki/data-viz_drug_consumption