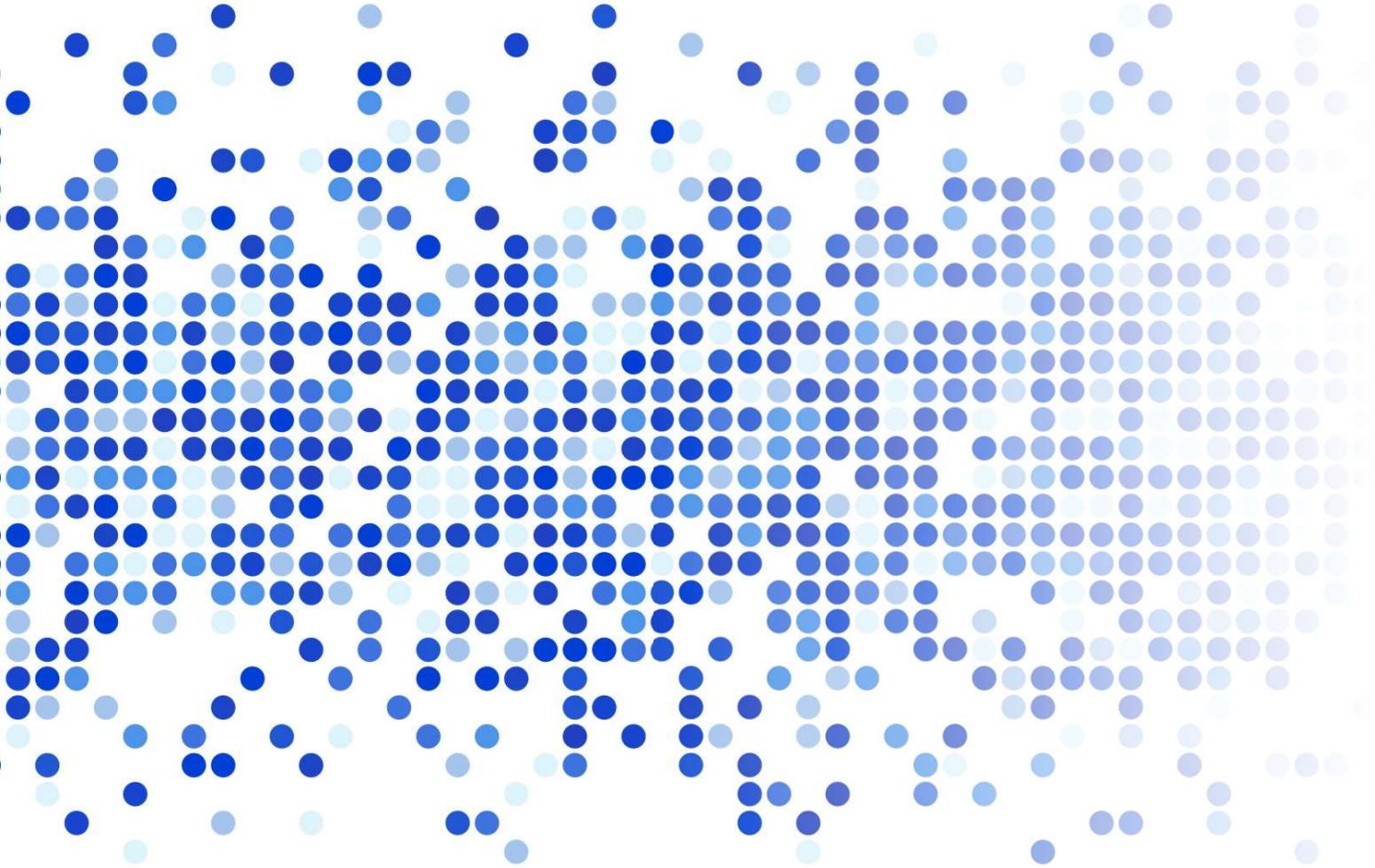

Python for Data Analysis Project

Report for the project containing the different steps of thinking and the explanation of the code I've been doing.

Index

Dataset information

First, we are introduced to the dataset. It is a dataset about drug consumption quantification. It has 1885 rows for 12 features such as :

- The first columns correspond to general information about participants (Age, gender, education level, country and ethnicity.
- Then comes a series of number that correspond to personality measurements
- And the last columns represent the last time the participant used one of the 18 legal and illegal drugs plus one fictious drug.

Data preprocessing

- The first thing that came to me is that the data set had no columns names and the categorical values such as the age or the country have all been replaced by values for each parameters to be normally distributed.

0	1	2	3	4	5	6	7	8	9	10	...	22	23	24	25	26	27	28	29	30
1	0.49788	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	...	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL2	CL0
2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	-0.14277	...	CL4	CL0	CL2	CL0	CL2	CL3	CL0	CL4	CL0
3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	...	CL0	CL0	CL0	CL0	CL0	CL0	CL1	CL0	CL0
4	-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	0.58489	...	CL0	CL0	CL2	CL0	CL0	CL0	CL0	CL2	CL0
5	0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	1.30612	...	CL1	CL0	CL0	CL1	CL0	CL0	CL2	CL2	CL0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1884	-0.95197	0.48246	-0.61113	-0.57009	-0.31685	-1.19430	1.74091	1.88511	0.76096	-1.13788	...	CL0	CL0	CL0	CL3	CL3	CL0	CL0	CL0	CL0
1885	-0.95197	-0.48246	-0.61113	-0.57009	-0.31685	-0.24649	1.74091	0.58331	0.76096	-1.51840	...	CL2	CL0	CL0	CL3	CL5	CL4	CL4	CL5	CL0
1886	-0.07854	0.48246	0.45468	-0.57009	-0.31685	1.13281	-1.37639	-1.27553	-1.77200	-1.38502	...	CL4	CL0	CL2	CL0	CL2	CL0	CL2	CL6	CL0
1887	-0.95197	0.48246	-0.61113	-0.57009	-0.31685	0.91093	-1.92173	0.29338	-1.62090	-2.57309	...	CL3	CL0	CL0	CL3	CL3	CL0	CL3	CL4	CL0
1888	-0.95197	-0.48246	-0.61113	0.21128	-0.31685	-0.46725	2.12700	1.65653	1.11406	0.41594	...	CL3	CL0	CL0	CL3	CL3	CL0	CL3	CL6	CL0

Naming columns

So, I decided to load the data set and this time put columns names explicitly from the

	Age	Gender	Education level	Country	Ethnicity	NScore	EScore	OScore	AScore	CScore	...	Ecstasy	Heroin	Ketamine	Legal highs	LSD	Meth
1	0.49788	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	...	CL0	CL0	CL0	CL0	CL0	
2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	-0.14277	...	CL4	CL0	CL2	CL0	CL2	
3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	...	CL0	CL0	CL0	CL0	CL0	
4	-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	0.58489	...	CL0	CL0	CL2	CL0	CL0	
5	0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	1.30612	...	CL1	CL0	CL0	CL1	CL0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1884	-0.95197	0.48246	-0.61113	-0.57009	-0.31685	-1.19430	1.74091	1.88511	0.76096	-1.13788	...	CL0	CL0	CL0	CL3	CL3	
1885	-0.95197	-0.48246	-0.61113	-0.57009	-0.31685	-0.24649	1.74091	0.58331	0.76096	-1.51840	...	CL2	CL0	CL0	CL3	CL5	
1886	-0.07854	0.48246	0.45468	-0.57009	-0.31685	1.13281	-1.37639	-1.27553	-1.77200	-1.38502	...	CL4	CL0	CL2	CL0	CL2	
1887	-0.95197	0.48246	-0.61113	-0.57009	-0.31685	0.91093	-1.92173	0.29338	-1.62090	-2.57309	...	CL3	CL0	CL0	CL3	CL3	
1888	-0.95197	-0.48246	-0.61113	0.21128	-0.31685	-0.46725	2.12700	1.65653	1.11406	0.41594	...	CL3	CL0	CL0	CL3	CL3	

Mapping value to a category

- Then, thanks to the map function in pandas and lambda functions I created, I replaced every value by its categorical representation.

For example in the age section :

- -0.95197 represent the category 18-24
- -0.07854 represent the category 25-34
- 0.49788 represent the category 35-44
- 1.09449 represent the category 45-54
- 1.82213 represent the category 55-64
- 2.59171 represent the category 65+

And I transformed the Age, Country, Education Level and Gender columns by doing this.

Changing the dataset

- I considered that the education level feature was divided in too many different categories (9) so I decided to merge some of them to drop to 5 (group all categories that left school without any for example).
- Then comes the ethnicity column. Since we are in France, and this type of feature is forbidden, I decided to drop this column even though it may have been helpful in the future.

Final look of the dataset

This is much neater

	Age	Gender	Education level	Country	NScore	EScore	OScore	AScore	CScore	Impulsive	...	Ecstasy	Heroin	Ketamine	Legal highs	LSD	Methadc
1	35-44	F	Professional certificate	UK	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	...	CL0	CL0	CL0	CL0	CL0	C
2	25-34	H	Doctorate degree	UK	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	...	CL4	CL0	CL2	CL0	CL2	C
3	35-44	H	Professional certificate	UK	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	...	CL0	CL0	CL0	CL0	CL0	C
4	18-24	F	Masters degree	UK	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	...	CL0	CL0	CL2	CL0	CL0	C
5	35-44	F	Doctorate degree	UK	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	...	CL1	CL0	CL0	CL1	CL0	C
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1884	18-24	F	No degree	USA	-1.19430	1.74091	1.88511	0.76096	-1.13788	0.88113	...	CL0	CL0	CL0	CL3	CL3	C
1885	18-24	H	No degree	USA	-0.24649	1.74091	0.58331	0.76096	-1.51840	0.88113	...	CL2	CL0	CL0	CL3	CL5	C
1886	25-34	F	University degree	USA	1.13281	-1.37639	-1.27553	-1.77200	-1.38502	0.52975	...	CL4	CL0	CL2	CL0	CL2	C
1887	18-24	F	No degree	USA	0.91093	-1.92173	0.29338	-1.62090	-2.57309	1.29221	...	CL3	CL0	CL0	CL3	CL3	C
1888	18-24	H	No degree	Ireland	-0.46725	2.12700	1.65653	1.11406	0.41594	0.88113	...	CL3	CL0	CL0	CL3	CL3	C

Personality measurements

- This is 7 columns that represents the subject personality. It is based on the subject's Big Five personality traits (which is well known in psychology) plus two other trait :

  - Neuroticism
  - Extraversion
  - Openness
  - Agreeableness
  - Conscientiousness
  - Impulsiveness
  - Sensation-seeing

- They are measured by NEO-FFI-R, BIS-11 and ImpSS.

Personality measurements

This time, I decided to not transform them because they represent a value in the test. Since we can't interpret those value, it's the same as leaving their numerical representation.

	NScore	EScore	OScore	AScore	CScore	Impulsive	Sensation seeing
1	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1.18084
2	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575
3	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0.40148
4	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084
5	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	-0.21575
...	...	...	...	...	...	...	...
1884	-1.19430	1.74091	1.88511	0.76096	-1.13788	0.88113	1.92173
1885	-0.24649	1.74091	0.58331	0.76096	-1.51840	0.88113	0.76540
1886	1.13281	-1.37639	-1.27553	-1.77200	-1.38502	0.52975	-0.52593
1887	0.91093	-1.92173	0.29338	-1.62090	-2.57309	1.29221	1.22470
1888	-0.46725	2.12700	1.65653	1.11406	0.41594	0.88113	1.22470

Drug consumption

- And finally we have the 18 columns plus one that represents each subject last consumption of a drug. Here is a list of all them :

- Alcohol	- Amphetamines	- Amyl nitrite	- Benzodiazepine
- Caffeine	- Cannabis	- Chocolate	- Cocaine
- Crack	- Ecstasy	- Heroin	- Ketamine
- Legal highs	- LSD	- Methadone	- Mushrooms
- Nicotine	- Semeron	- Volatile Substances	

Drug consumption

- We note the presence of Semeron : this is a fictious drug that has been introduced here to detect the over-claimers. We also note the presence of Chocolate or Caffeine which are considered as drug regarding they addictiveness.

- Each level in each drug represent the last time the subject used it :

- 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

Data visualisation

Thanks to this really complex dataset, we could have ask ourselves a lot of question regarding the data and it analysis. But we had to isolate a few of them and here are the three I decided to investigate :

- What is the repartition of drug consumer according to their education level and their gender ?
- Are our data representative of this country population ? If so what is the repartition of drug consumer per country ?
- Which drugs are the most addictive ?

Drug consumption per education and gender

- The plots produced can be found in the notebook.
- To ease the task, I decided to consider that if the subject consumed a drug in the last month, he was a what we could call a regular consumer.
- The plots show us that with every drug the category that is the more likely to be consumer is the one without any degree or diploma. Furthermore, in most of the cases, the men were more likely to be consumer than the women.

Data representativeness and consumer per country

- Same as before, the plots can be found in the notebook. And we used our regular consumers.
- If we consider our data to be representative, the country that has the most consumer for every drug would be the USA, with approximately 6 times more consumer than every other country.
- If we take as example the alcohol consumption for United Kingdom : with around 60 000 000 people drinking every month this would represents 92% of the UK citizens that drink which is way more than what we found doing research.

Drug addictiveness

- This time we decided to split the consumers into three groups :

- Regular consumer

- Once Consumer

- Never consumed

And decided to sorted the results by the descending number of regular consumers.

Top 10 most addictive drug among our dataset

	Chocolate	Caffeine	Alcohol	Nicotine	Cannabis	Benzodiazepine	Legal highs	Ecstasy	Amphetamines	Methadone
Regular consumer	1786	1764	1551	875	788	299	241	240	238	171
Once consumed	64	84	266	389	477	470	521	511	441	246
Not consumer	35	37	68	621	620	1116	1123	1134	1206	1468

- As we could have thought, the chocolate is the first here, probably because it is not considered as a drug in the common sense, so is Caffeine.
- Without any sort of surprise the first four are all legal drugs, and the two following are well known drug, the first one is considered by some as a recreational drug, when the second is a drug that can be consumed to heal depression. However we can see a big jump between the fifth and the sixth drug in number of consumers.

Machine Learning models

- Here we had to convert our features using One Hot Encoder.
- Instead of trying to predict all 18 drug consumption at the same time (with 7 results classes), that could make a mess and not only take longer to run but generate false results; I decided to focus on few drug and work especially with two : the Heroin and the Cannabis.
- While the heroin model worked well pretty fast (85% accuracy on first try), the Cannabis one struggle to get close to 50% accuracy. This is probably because as said before, this is seen as a recreational drug and so the profile of regular consumers is much more diverse than for Heroin
- This time the results can be found under the form of an API which instructions to run are in the README.md file.

Conclusion

- I think that here we did everything we had to : there was data pre-processing, data visualisation, data modelling and even data prediction.
- We saw that our dataset was not representative enough despite the number of subjects (1558 is pretty good) which is something that could maybe be corrected by applying a weight to each subject.
- Also, in spite of the range of features, we struggled to predict well some class for some drug consumption, which shows us that the human nature is way more complex than a simple Machine Learning Algorythm.