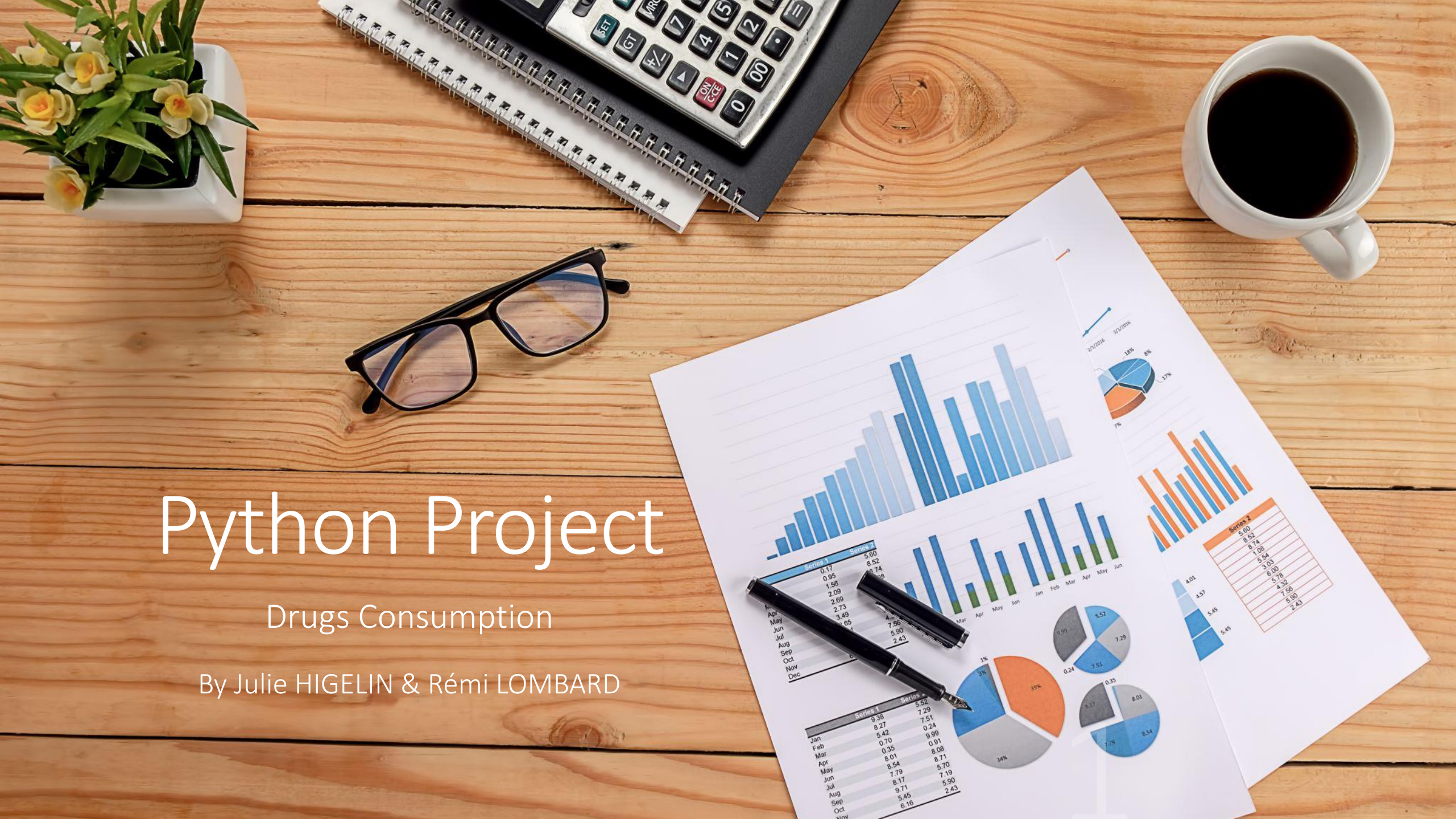
Python Project

Drugs Consumption

By Julie HIGELIN & Rémi LOMBARD



Summary

- Ins of the study
- Outs of the study
- Our thougths on the asked question
- How the problem fits in the context of the study

Ins of the study

Ins of the study

As in, we have a dataset of 1884 rows and 32 columns:

	1	0.49788	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	...	CL0.4	CL0.5	CL0.6	CL0.7	CL0.8	CL0.9	CL0.10	CL2.2	CL0.11	CL0.12
0	2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	...	CL4	CL0	CL2	CL0	CL2	CL3	CL0	CL4	CL0	CL0
1	3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	...	CL0	CL0	CL0	CL0	CL0	CL0	CL1	CL0	CL0	CL0
2	4	-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	...	CL0	CL0	CL2	CL0	CL0	CL0	CL0	CL2	CL0	CL0
3	5	0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	...	CL1	CL0	CL0	CL1	CL0	CL0	CL2	CL2	CL0	CL0
4	6	2.59171	0.48246	-1.22751	0.24923	-0.31685	-0.67825	-0.30033	-1.55521	2.03972	...	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0	CL0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1879	1884	-0.95197	0.48246	-0.61113	-0.57009	-0.31685	-1.19430	1.74091	1.88511	0.76096	...	CL0	CL0	CL0	CL3	CL3	CL0	CL0	CL0	CL0	CL5
1880	1885	-0.95197	-0.48246	-0.61113	-0.57009	-0.31685	-0.24649	1.74091	0.58331	0.76096	...	CL2	CL0	CL0	CL3	CL5	CL4	CL4	CL5	CL0	CL0
1881	1886	-0.07854	0.48246	0.45468	-0.57009	-0.31685	1.13281	-1.37639	-1.27553	-1.77200	...	CL4	CL0	CL2	CL0	CL2	CL0	CL2	CL6	CL0	CL0
1882	1887	-0.95197	0.48246	-0.61113	-0.57009	-0.31685	0.91093	-1.92173	0.29338	-1.62090	...	CL3	CL0	CL0	CL3	CL3	CL0	CL3	CL4	CL0	CL0
1883	1888	-0.95197	-0.48246	-0.61113	0.21128	-0.31685	-0.46725	2.12700	1.65653	1.11406	...	CL3	CL0	CL0	CL3	CL3	CL0	CL3	CL6	CL0	CL2

1884 rows × 32 columns

Ins of the study

The 32 columns are as follows:

- Age
- Education_Level
- Ethnicity
- Extraversion
- Ascore
- Impulsive
- Alcohol
- Caffeine
- Chocolat
- Crack
- Heroin
- Legalh
- Meth
- Nicotine
- VSA
- Gender
- Country
- Nscore
- Oscore
- Cscore
- SS
- Amphet
- Benzos
- Cannabis
- Cocaïne
- Ecstasy
- Ketamine
- LSD
- Mushrooms
- Semer
- ID

	ID	Age	Gender	Education_Level	Country	Ethnicity	Nscore	Extraversion	Oscore	Ascore	...	Ecstasy	Heroin	Ketamine	Legalh	LSD	Meth	Mushrooms	Nicotine	Semer	VSA
0	2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	...	CL4	CL0	CL2	CL0	CL2	CL3	CL0	CL4	CL0	CL0
1	3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	...	CL0	CL0	CL0	CL0	CL0	CL0	CL1	CL0	CL0	CL0
2	4	-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	...	CL0	CL0	CL2	CL0	CL0	CL0	CL0	CL2	CL0	CL0
3	5	0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	...	CL1	CL0	CL0	CL1	CL0	CL0	CL2	CL2	CL0	CL0
4	6	2.59171	0.48246	-1.22751	0.24923	-0.31685	-0.67825	-0.30033	-1.55521	2.03972	...	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL6	CL0	CL0

Ins of the study

Features	For Visualization	For Modelization
Age	18-24	18
	25-34	25
	35-44	35
	45-54	45
	55-64	55
	65+	65
Gender	0 (Male)	0
	1 (Female)	1
Education Level	Left school before 16 years	0
	Left school at 16 years	1
	Left school at 17 years	2
	Left school at 18 years	3
	Some college or university, no certificate or degree	4
	Professional certificate/ diploma university	5 6
	Masters degree	7
	Doctorate degree / PHD	8
Country	Australia	0
	Canada	1
	New Zealand	2
	Republic of Ireland	3
	UK	4
	USA	5
	Other	6
Ethnicity	Asian	0
	Black	1
	Mixed-Black/Asian	2
	Mixed-White/Asian	3
	Mixed-White/Black	4
	White	6
	Other	5

Features	For Visualization	For Modelization
Nscore (Neuroticism)	Scores : from 12 to 61	idem
Extraversion	Scores : 16, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 58, 59	Idem
Oscore (Openness to experience)	Scores : 24, 26, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60	Idem
Ascore (Agreeableness)	Scores : 12, 16, 18, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60	Idem
Cscore (Conscientiousness)	Scores : 17, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 59	Idem
Impulsivity	Scores : 20, 276, 307, 355, 257, 216, 195, 148, 104, 7	Idem
SS (Sensation seeking)	Scores : 71 , 87, 132, 169, 211,223,219,249,211,210,103	Idem
Drugs : Alcohol, Amphet, Amyl, Benzos, Caffeine, Cannabis, Chocolat, Cocaine, Crack, Ecstasy, Heroin, Ketamine, Legalh, Meth, Mushrooms, Nicotine, Semer (fictitious drug), VSA	0 (Never Used)	0
	1 (Used over a Decade Ago)	1
	2 (Used in Last Decade)	2
	3 (Used in Last Year)	3
	4 (Used in Last Month)	4
	5 (Used in Last Week)	5
	6 (Last Day)	6
ID	Depends on the line	Idem

Ins of the study

We removed everyone who said he had used Semer. Indeed the Semer is a false drug, so all the people claiming to consume Semer are liars. 8 people said they had consumed semer, so we deleted their line. Finally, we have deleted the semer column.

We also deleted the ID column because an identifier is unique, so this column will not help us make predictions

At the end of the cleaning, we have 1876 rows and 30 columns

	Age	Gender	Education_Level	Country	Ethnicity	Nscore	Extraversion	Oscore	Ascore	Cscore	...	Crack	Ecstasy	Heroin	Ketamine	Legalh	LSD	Meth	Mushr
0	25-34	0	PHD	UK	White	29.0	52	55	48	41	...	0	4	0	2	0	2	3	
1	35-44	0	college with degree	UK	White	31.0	45	40	32	34	...	0	0	0	0	0	0	0	
2	18-24	1	masters	UK	White	34.0	34	46	47	46	...	0	0	0	2	0	0	0	
3	35-44	1	PHD	UK	White	43.0	28	43	41	50	...	0	1	0	0	1	0	0	
4	65+	1	quit at 18	Canada	White	29.0	38	35	55	52	...	0	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1879	18-24	1	college without degree	USA	White	25.0	51	57	48	33	...	0	0	0	0	3	3	0	
1880	18-24	0	college without degree	USA	White	33.0	51	50	48	30	...	0	2	0	0	3	5	4	
1881	25-34	1	university	USA	White	47.0	30	37	31	31	...	0	4	0	2	0	2	0	
1882	18-24	1	college without degree	USA	White	45.0	26	48	32	22	...	0	3	0	0	3	3	0	
1883	18-24	0	college without degree	Ireland	White	31.0	53	56	50	45	...	0	3	0	0	3	3	0	

1876 rows × 30 columns

Outs of the study

Outs of the study

As outs, we predicted for an individual all the mastiffs that they are likely to consume. In the columns we find the quantity of drugs consumed, and in the row, the details of the drugs consumed.

Let's take an example :

For a given dataset, we are going to focus on the first row, so the first individual. The first individual is a 25–35-year-old white American male with a doctorate who scored 29, 52, 55, 48, 41, 307, 223 on the personality test. The model predicts that this person will probably take 5 drugs: Alcohol, Caffeine, Cannabis, Chocolate and Nicotine.

More generally, a prediction is made in the following format:

Ins:

	Age	Gender	Education_Level	Country	Ethnicity	Nscore	Extraversion	Oscore	Ascore	Cscore	Impulsive	SS
0	25	0	8	5	6	29.0	52	55	48	41	307	223
1	35	0	5	5	6	31.0	45	40	32	34	276	249
2	18	1	7	5	6	34.0	34	46	47	46	276	132
3	35	1	8	5	6	43.0	28	43	41	50	355	223
4	65	1	3	1	6	29.0	38	35	55	52	276	87

Outs:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	Alcohol	Caffeine	Cannabis	Chocolat	Nicotine	-	-	-	-	-	-	-	-	-
1	Alcohol	Caffeine	Chocolat	-	-	-	-	-	-	-	-	-	-	-
2	Alcohol	Caffeine	Chocolat	-	-	-	-	-	-	-	-	-	-	-
3	Alcohol	Caffeine	Chocolat	-	-	-	-	-	-	-	-	-	-	-
4	Alcohol	Caffeine	Chocolat	Nicotine	-	-	-	-	-	-	-	-	-	-
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1853	Alcohol	Amphet	Caffeine	Cannabis	Chocolat	Ecstasy	Legalh	LSD	Mushrooms	Nicotine	-	-	-	-
1854	Alcohol	Caffeine	Cannabis	Chocolat	Ecstasy	Legalh	LSD	Mushrooms	Nicotine	-	-	-	-	-
1855	Alcohol	Amphet	Benzos	Caffeine	Cannabis	Chocolat	Cocaine	Nicotine	-	-	-	-	-	-
1856	Alcohol	Amphet	Benzos	Caffeine	Cannabis	Chocolat	Cocaine	Ecstasy	Legalh	LSD	Meth	Mushrooms	Nicotine	-
1857	Alcohol	Amphet	Caffeine	Cannabis	Chocolat	Ecstasy	Legalh	Mushrooms	Nicotine	-	-	-	-	-

Outs of the study

To make the last prediction, we tested several models.

Model
Random Forest
Decision Tree
Random Forest Grid Random
Random Forest Grid with Cross Val
Knn Classification
Logistic Regression

Random Forest with grid search offering the best results with over 86.3% accuracy, it is the one we used for the final prediction.

Score Test	Model
86.352	Random Forest Grid with Cross Val
86.210	Random Forest
85.890	Logistic Regression

Our thougths on the asked question

Our thoughts on the asked question

There was a lot of research capability with this dataset. We were offered the following problems:

- 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.

Our thoughts on the asked question

We decided to respond to as many problems as possible by posing an even more general problem :

"What drugs is an individual likely to take?“

This question involves a search by classification for each drug and also the transformation of the dataset to make it a binary search (an individual takes or doesn’t take a drug)