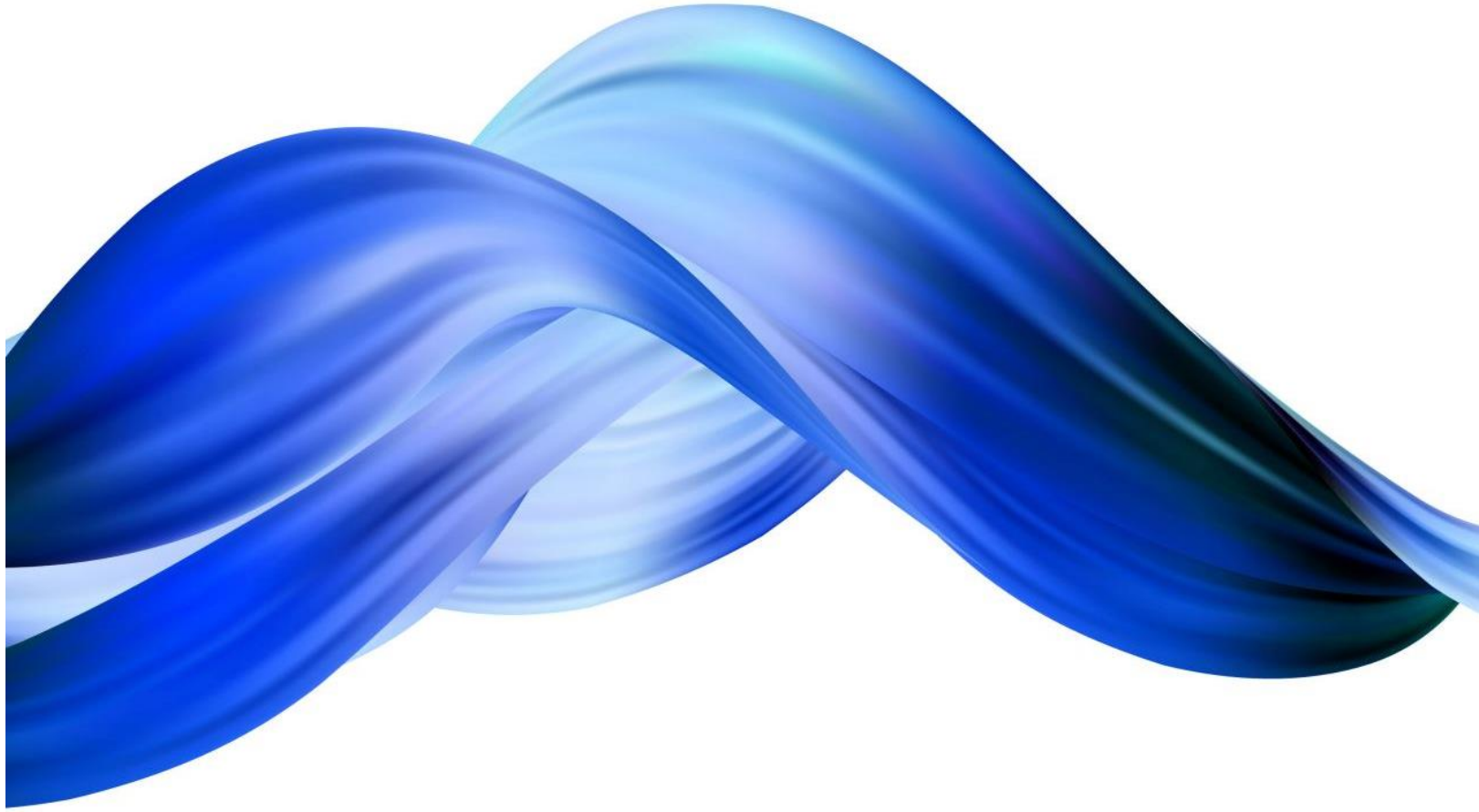
Drug Consumption analyzed with Machine Learning



By Othilie Cassin, Elias Bouasria and Guillaume de Trentinian

Context

- Study made with 1885 respondents on October 2016
- Explication about ins and outs of the study :
- 12 attributes are known : age, gender, country of residence etc.
- There are 18 legal and illegal drugs studied in this dataset. For each drug they have to select one of the answers: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day

Issues

- What are the correlations between one's drug consumption and their personality ?
- Evaluation of the drug consumption of (considering your attributes)

Data cleaning

Semer is a control variable:

we remove all rows where semer value is not 'Never used’

Choice to remove : « Ethnicity » from our dataset

We remove "Country" column

Classification of drug usage :

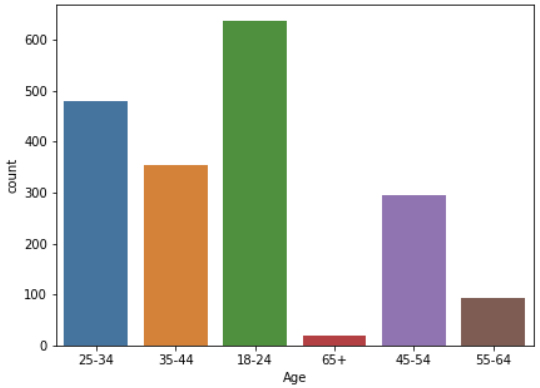
CL0 : Never Used

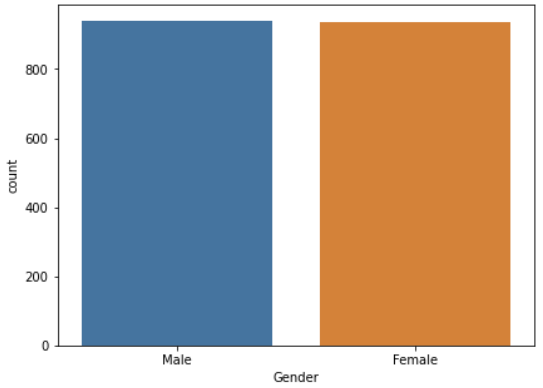
CL1-3 : Former User

CL4-6 : Current User

The study was made with special care to gender equality

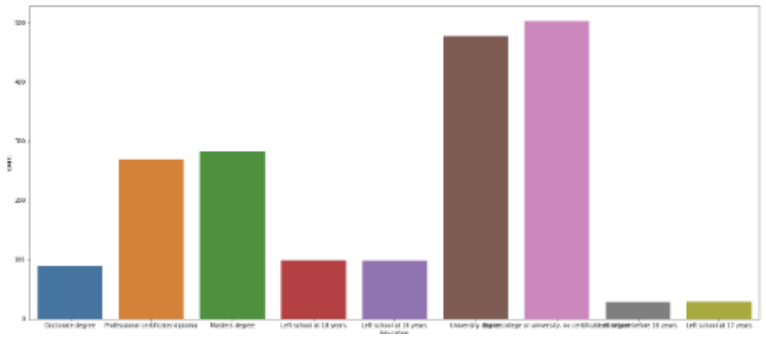
Age

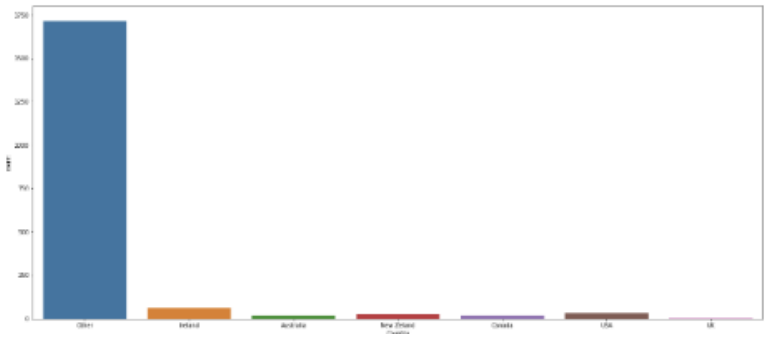




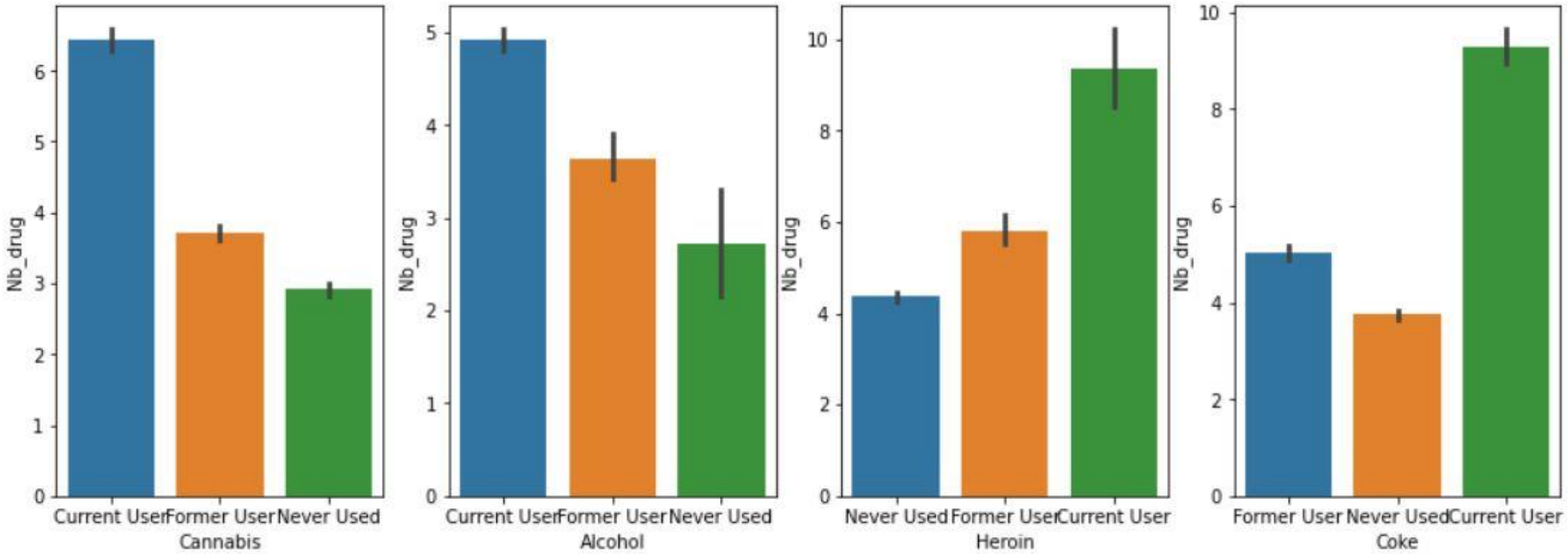
Gender

Education



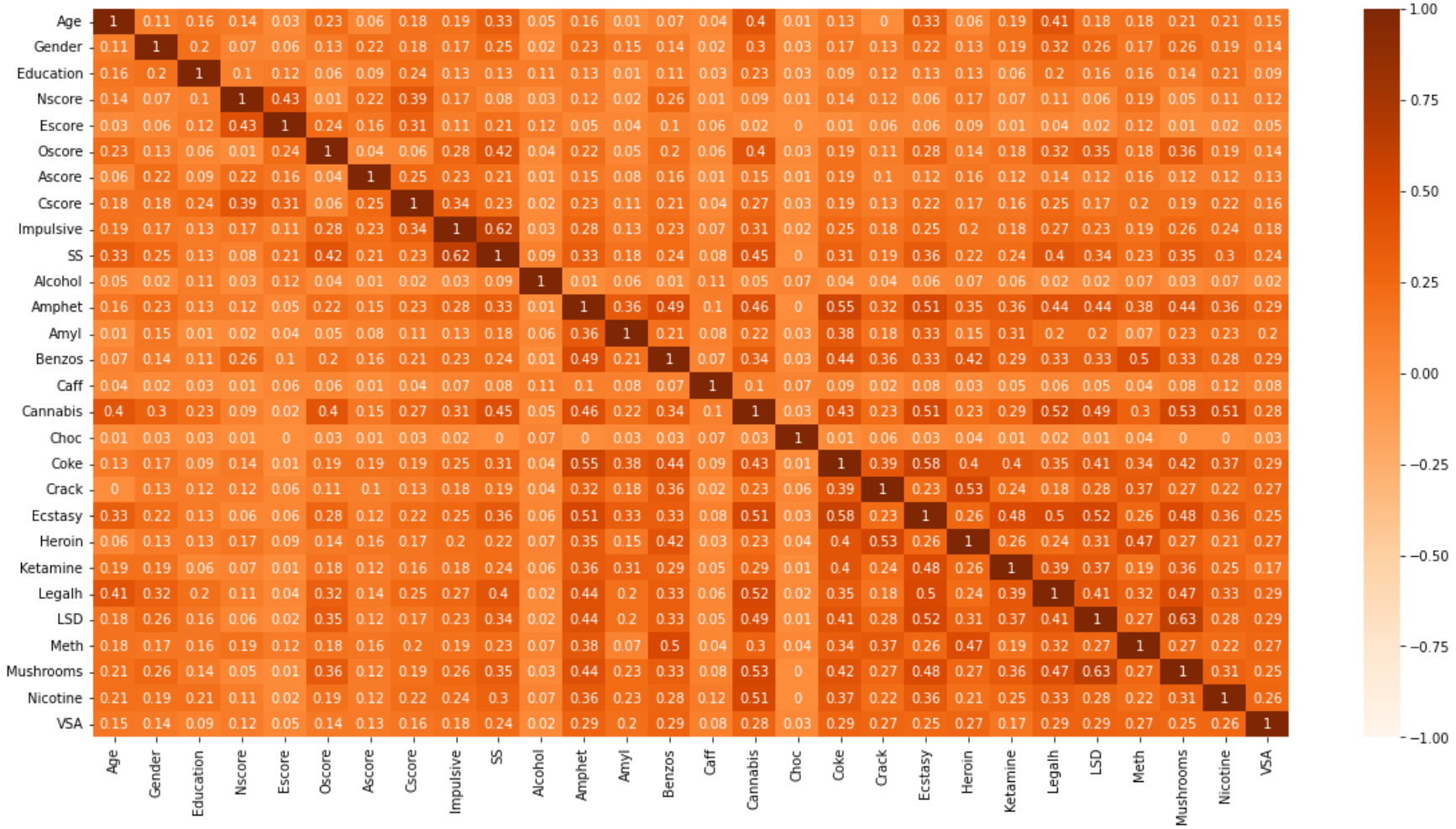


Country

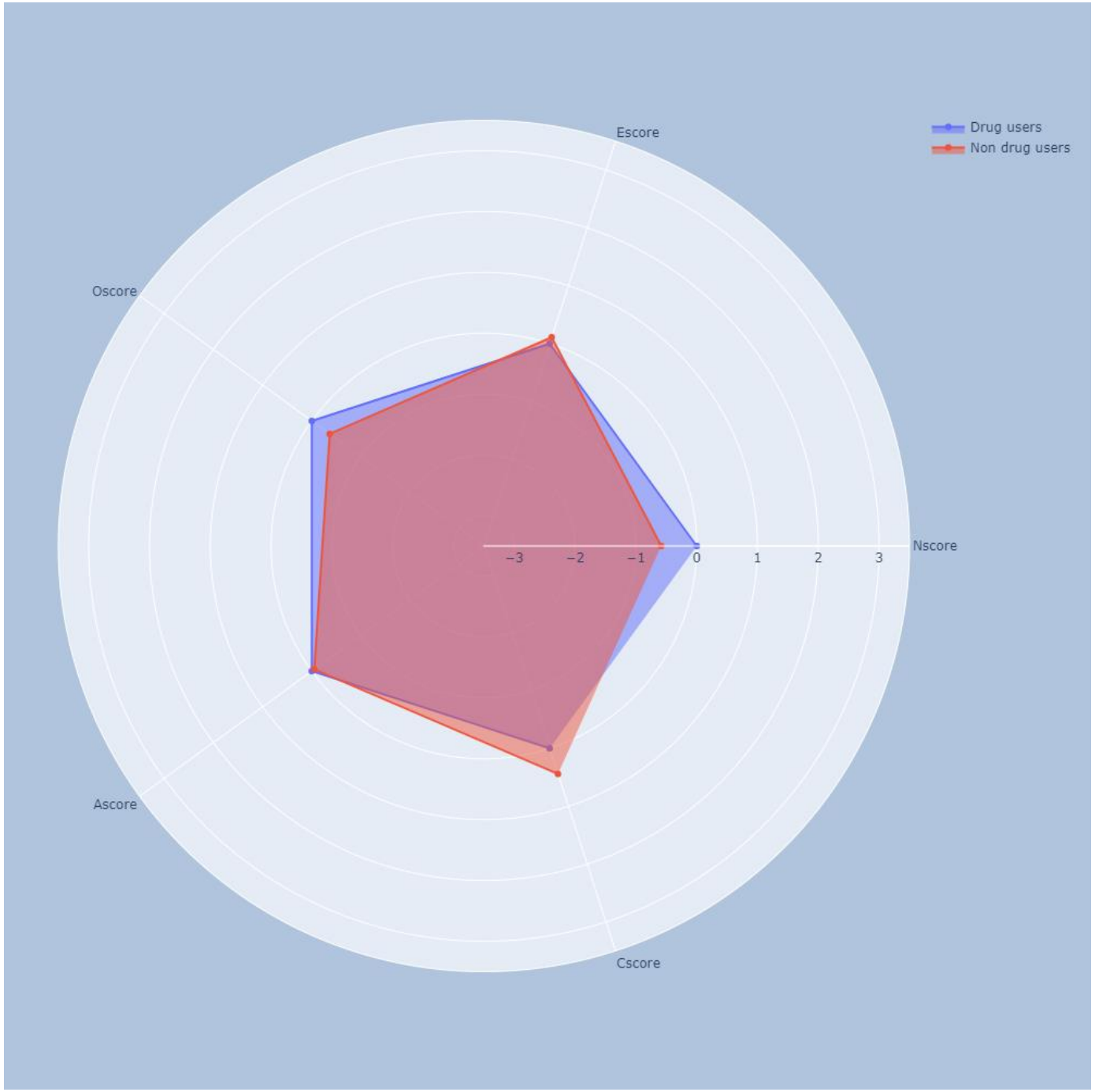


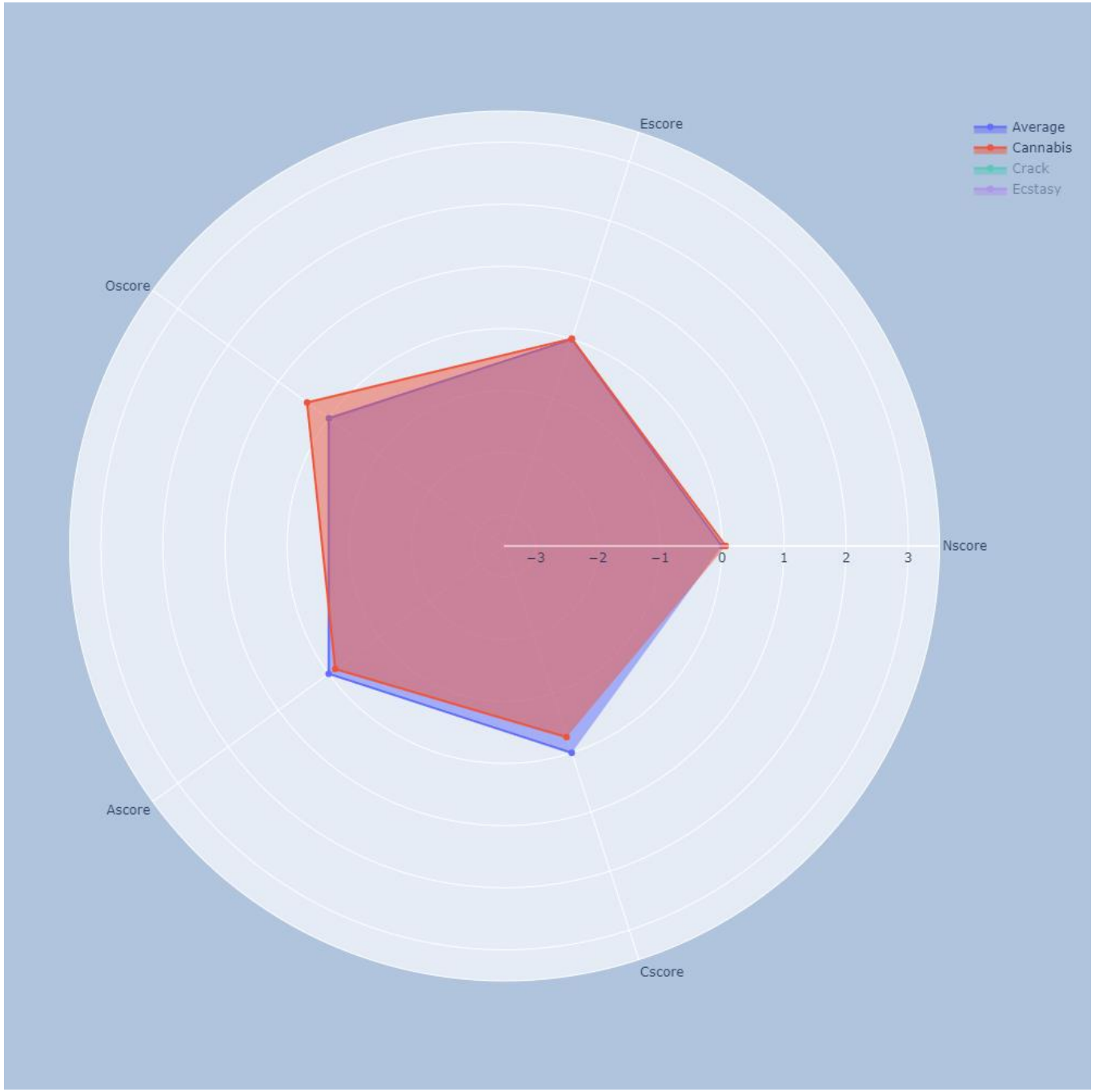
We created a new variable : "Number of drugs" to show if people who use some drugs tend to be using other drugs as well.

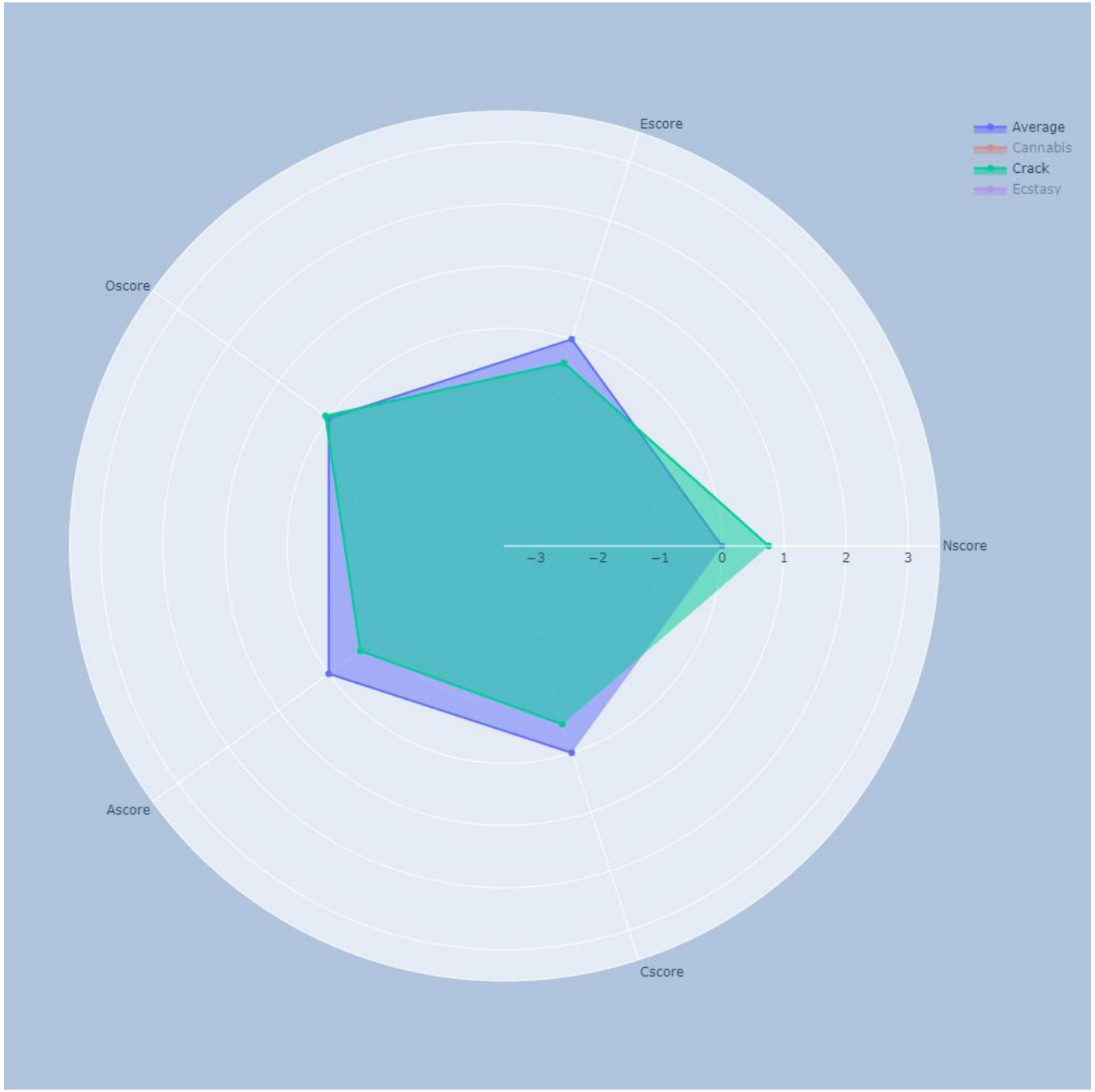
Correlation matrix

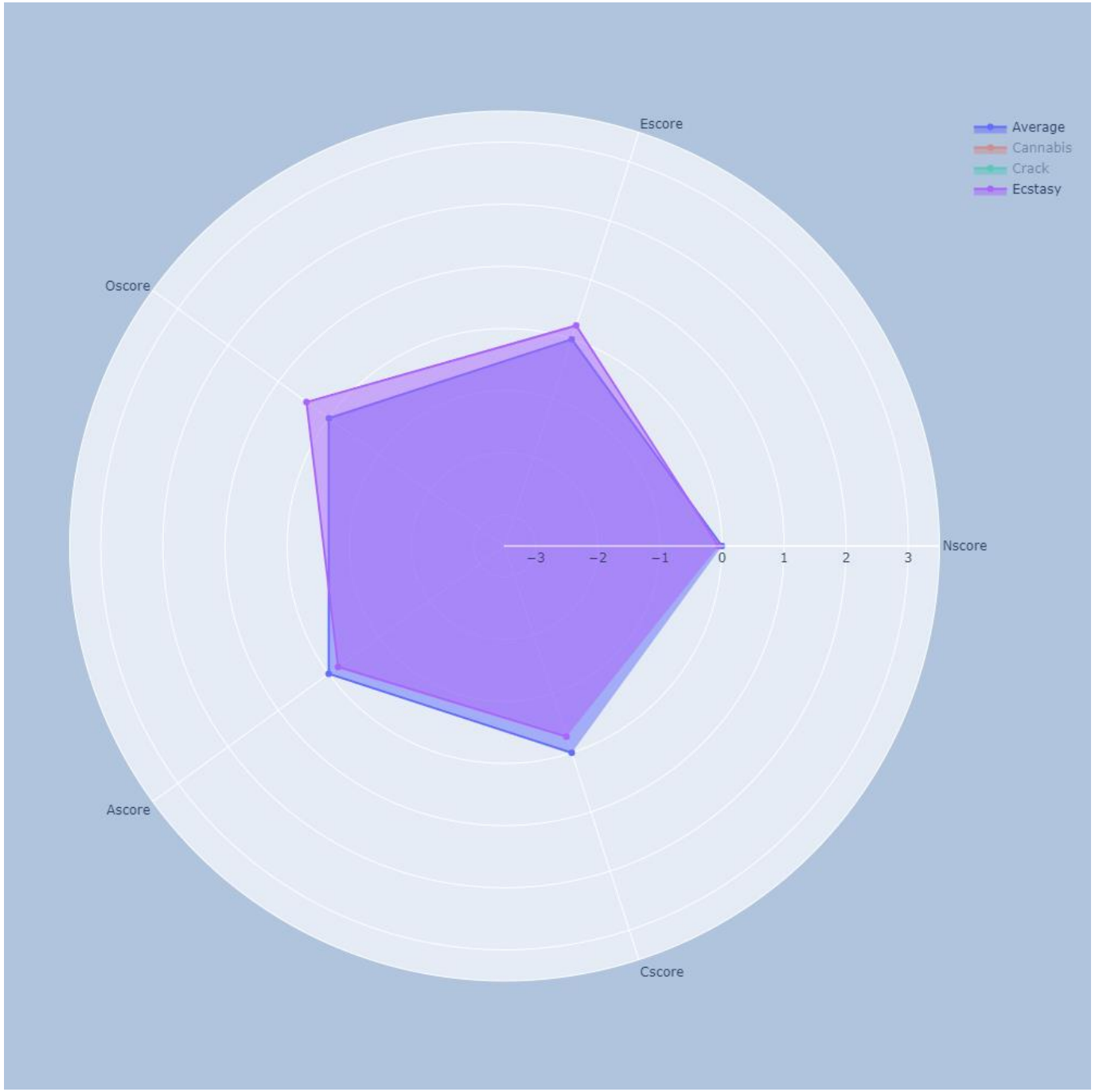


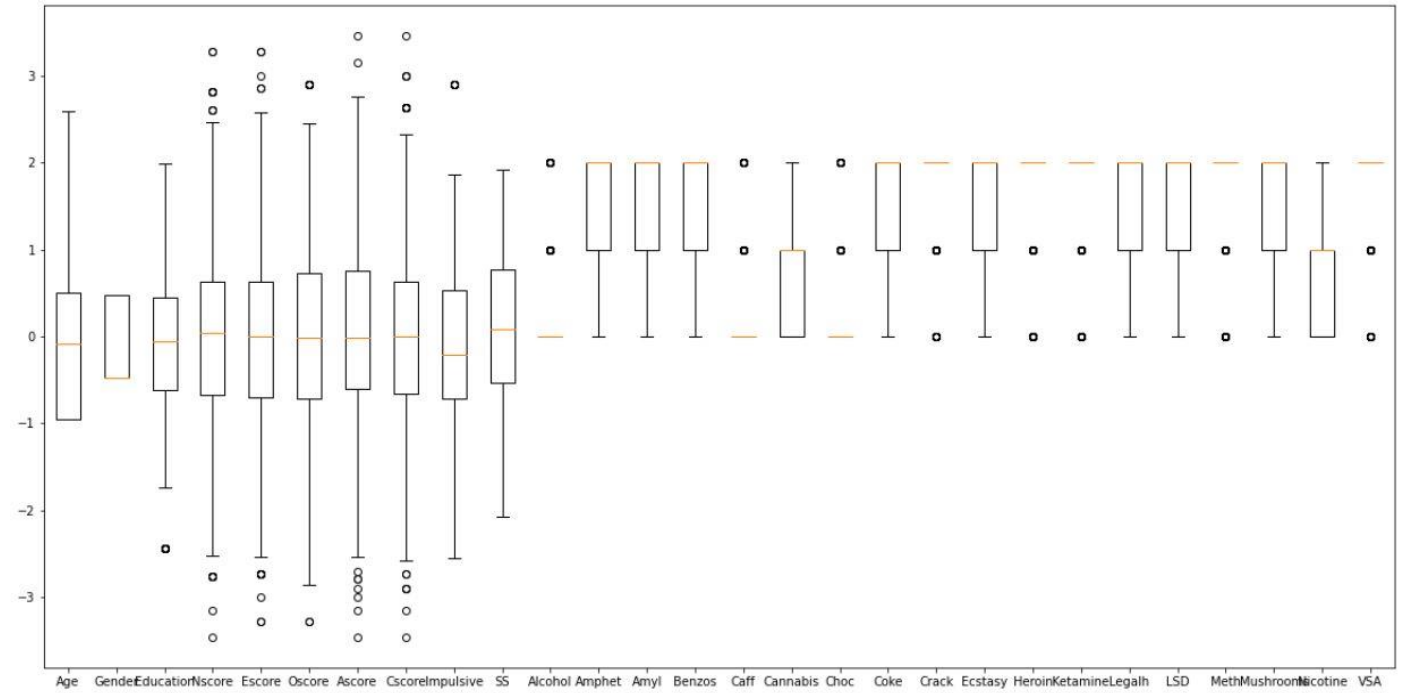
Personnality



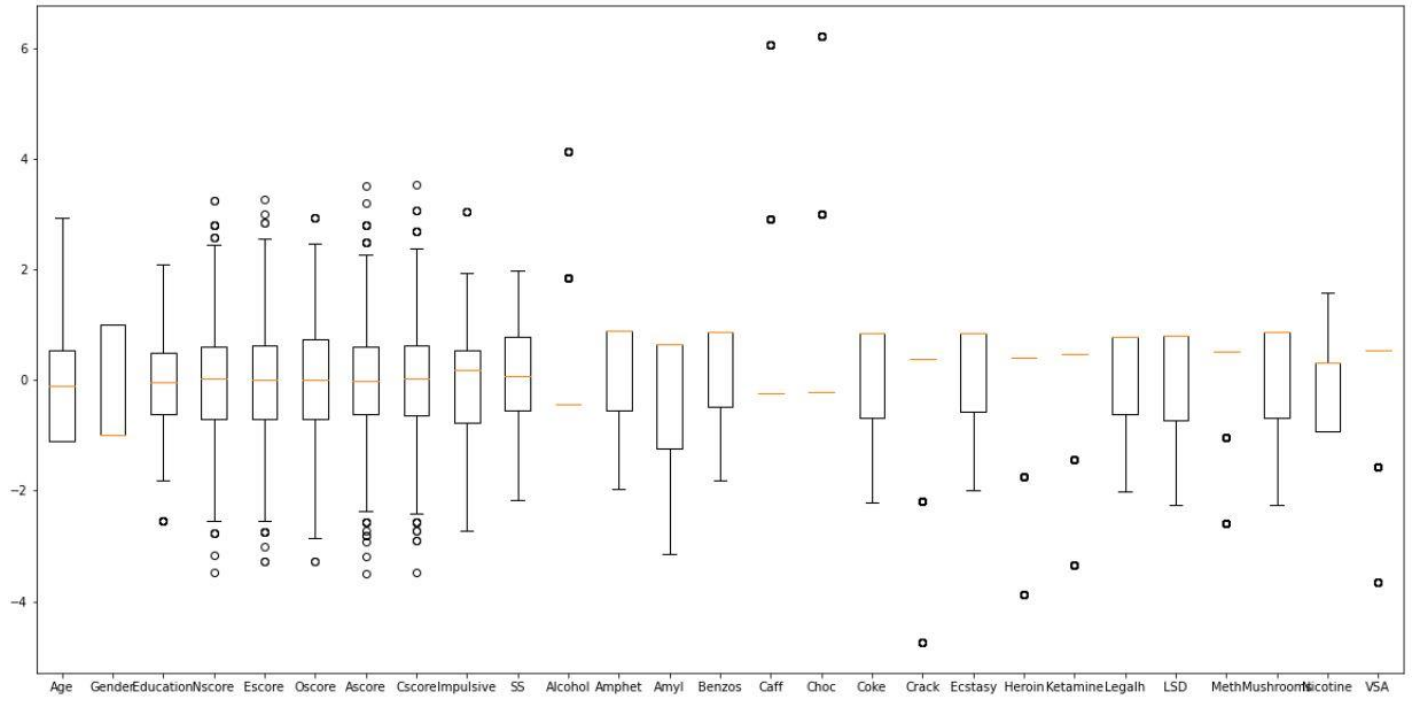

Escore
Oscore
Ascore
Cscore
Nscore
−3
−2
−1
0
1
2
3
Average
Cannabis
Crack
Ecstasy


Escore
Oscore
Ascore
Cscore
Nscore
−3
−2
−1
0
1
2
3
Average
Cannabis
Crack
Ecstasy


Escore
Oscore
Ascore
Cscore
Nscore
−3
−2
−1
0
1
2
3
Average
Cannabis
Crack
Ecstasy



Distribution of each feature before standardization
Although the first 12 columns have already been standardized by the UCI organization



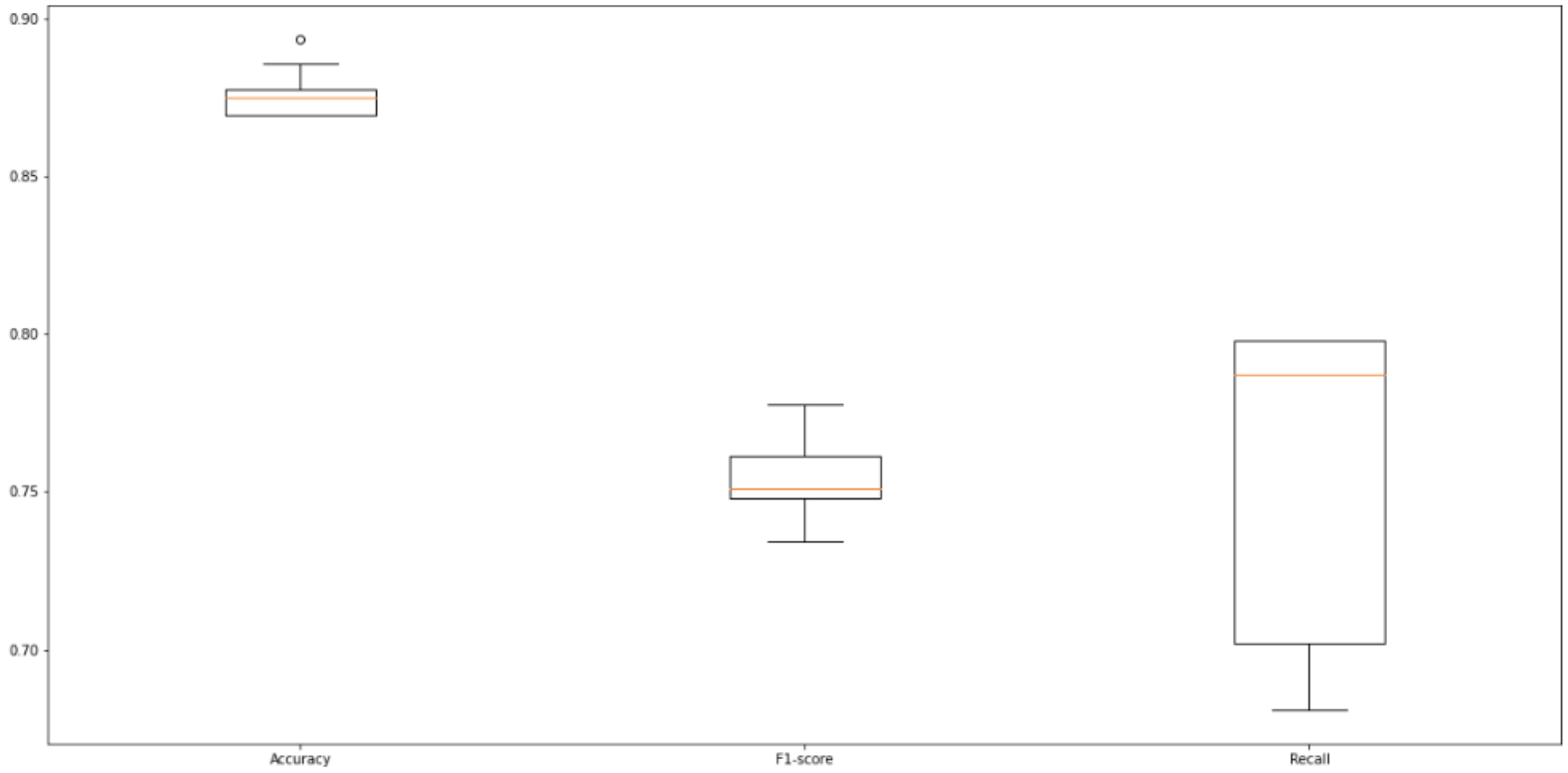
Distribution of each feature after standardization; the data is centered around 0 and grouped evenly for the first variables, the rest comes from unevenly distributed categorical variables

We tried to predict if cannabis has ever been used by the participant based on their personality traits and information

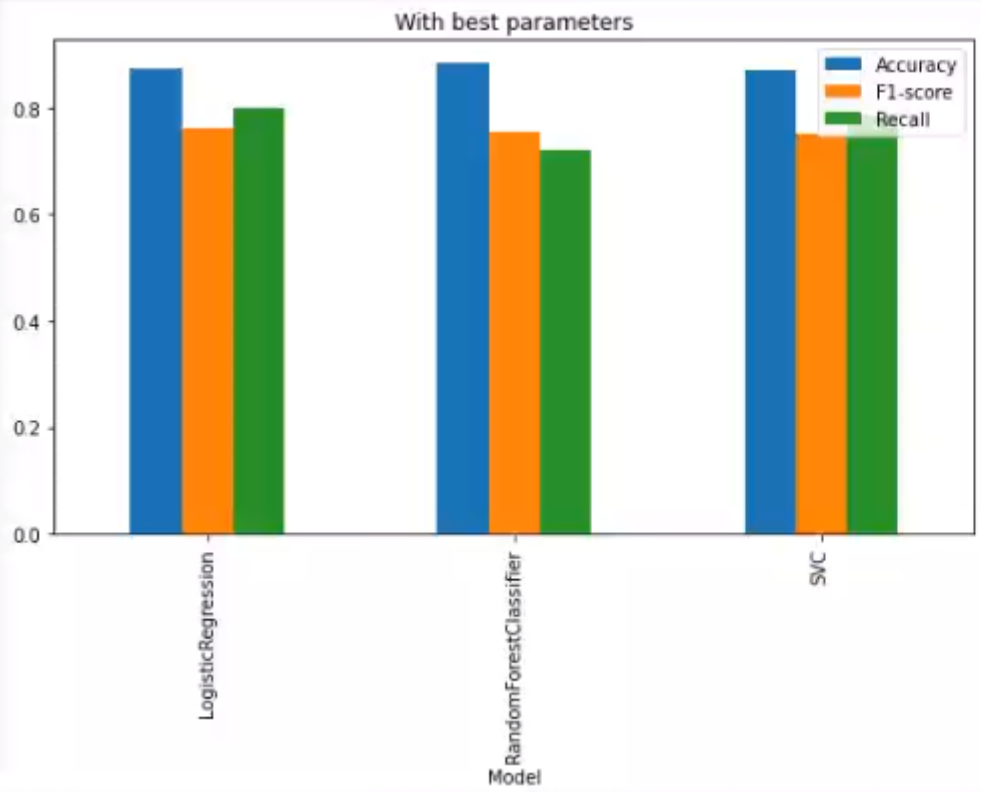
Creation of the column : Cannabis use (set on 0 if the participant never used cannabis and set on 1 otherwise) -> Our target for predictions

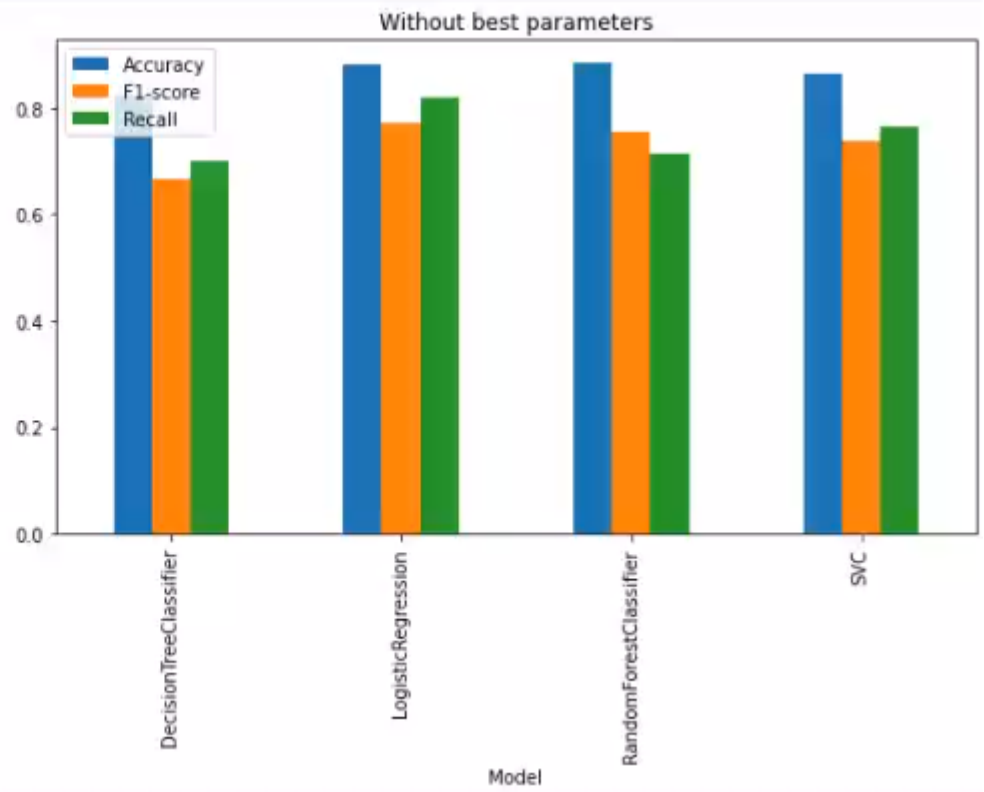
Accuracy of our regression models on the standardized data

	Model	Accuracy	F1-score	Recall
0	RandomForestClassifier	0.875000	0.734463	0.691489
1	RandomForestClassifier	0.882979	0.747126	0.691489
2	RandomForestClassifier	0.877660	0.738636	0.691489
3	RandomForestClassifier	0.882979	0.747126	0.691489
4	RandomForestClassifier	0.885638	0.754286	0.702128
5	RandomForestClassifier	0.880319	0.745763	0.702128
6	RandomForestClassifier	0.877660	0.741573	0.702128
7	RandomForestClassifier	0.893617	0.777778	0.744681
8	RandomForestClassifier	0.880319	0.745763	0.702128
9	RandomForestClassifier	0.880319	0.739884	0.680851
10	LogisticRegression	0.875000	0.761421	0.797872
11	LogisticRegression	0.875000	0.761421	0.797872
12	LogisticRegression	0.875000	0.761421	0.797872
13	LogisticRegression	0.875000	0.761421	0.797872
14	LogisticRegression	0.875000	0.761421	0.797872
15	LogisticRegression	0.875000	0.761421	0.797872
16	LogisticRegression	0.875000	0.761421	0.797872
17	LogisticRegression	0.875000	0.761421	0.797872
18	LogisticRegression	0.875000	0.761421	0.797872
19	LogisticRegression	0.875000	0.761421	0.797872
20	SVC	0.869681	0.751269	0.787234
21	SVC	0.869681	0.751269	0.787234
22	SVC	0.869681	0.751269	0.787234
23	SVC	0.869681	0.751269	0.787234
24	SVC	0.869681	0.751269	0.787234
25	SVC	0.869681	0.751269	0.787234
26	SVC	0.869681	0.751269	0.787234
27	SVC	0.869681	0.751269	0.787234
28	SVC	0.869681	0.751269	0.787234
29	SVC	0.869681	0.751269	0.787234

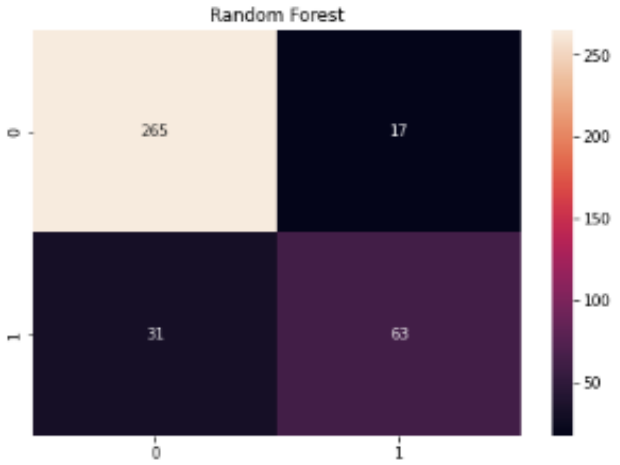


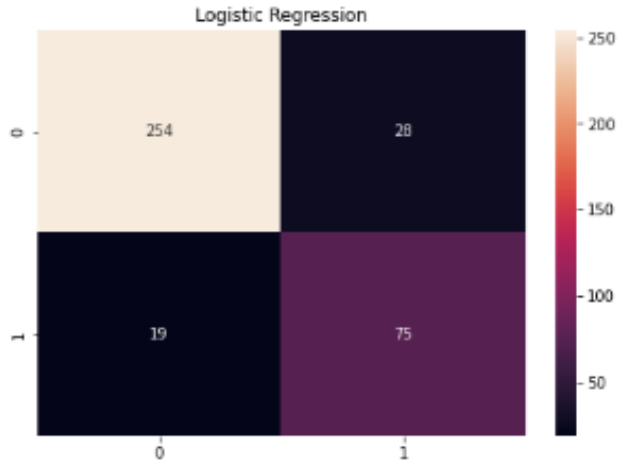
Gridsearch

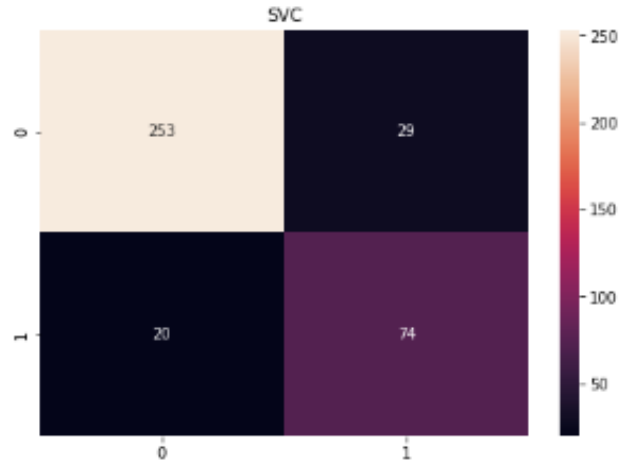

With best parameters
Accuracy
F1-score
Recall
0.0
0.2
0.4
0.6
0.8
LogisticRegression
RandomForestClassifier
SVC
Model


Without best parameters
Accuracy
F1-score
Recall
0.0
0.2
0.4
0.6
0.8
DecisionTreeClassifier
LogisticRegression
RandomForestClassifier
SVC
Model

Confusion matrixes







Conclusion and opening

Determine which in which educational paths alcohol consumption is the highest : can help to target a sensibilization campaign

Thanks for your attention !

Clickable link to our GitHub