

Final Project
Python for Data Analysis

Drug Consumption Analysis & Predictions

Cyprien NICOLAY - Timothé VITAL - Anna ZENOU
DIA 5

Summary



Drug Consumption Dataset presentation

Main information about the dataset and its organization



Data Pre-Processing

How we processed the dataset to use it efficiently



Data Visualizations

Visualizations of the dataset's principal information and the links between the variables and the target



Data Modeling

Different algorithms applied to the dataset

Drug Consumption Dataset

Link : <https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29#>

 1885 responses

	ID	Age	Genre	Education	Pays	Ethnicité	Neuroticisme	Extraversion	Ouverture à l'expérience	Agréabilité	Sérieux	Impulsivité	Recherche de sensation
0	1	0.49788	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1.18084
1	2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575
2	3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0.40148
3	4	-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084
4	5	0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	-0.21575

Alcool	Amphetamine	Amyl nitrite	Benzodiazepine	Caffeine	Cannabis	Chocolat	Cocaine	Crack	Ecstasy	Heroin	Ketamine	Substance psychoactive	LSD	Methadone	Champignon hallucinogène	Nicotine	Semeron	Substance Volatile
CL5	CL2	CL0	CL2	CL6	CL0	CL5	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL2	CL0	CL0
CL5	CL2	CL2	CL0	CL6	CL4	CL6	CL3	CL0	CL4	CL0	CL2	CL0	CL2	CL3	CL0	CL4	CL0	CL0
CL6	CL0	CL0	CL0	CL6	CL3	CL4	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL0	CL1	CL0	CL0	CL0
CL4	CL0	CL0	CL3	CL5	CL2	CL4	CL2	CL0	CL0	CL0	CL2	CL0	CL0	CL0	CL0	CL2	CL0	CL0
CL4	CL1	CL1	CL0	CL6	CL3	CL6	CL0	CL0	CL1	CL0	CL0	CL1	CL0	CL0	CL2	CL2	CL0	CL0

Drug Consumption Dataset

■ 5 demographic features :

- Age
- Gender
- Level of education
- Country
- Ethnicity

■ 7 personality features :

- Neuroticism
- Extraversion
- Openness to experience
- Agreeableness
- Conscientiousness
- Impulsiveness
- Sensation seeking

All input attributes are originally categorical and are quantified.
After quantification, values of all input features can be considered as real-valued.

Drug Consumption Dataset

18 drugs :

- Alcohol
- Amphetamines
- Amyl nitrite
- Benzodiazepine
- Caffeine
- Chocolate
- Cocaïne
- Crack
- Ecstasy
- Heroin
- Ketamine
- Legal highs
- LSD
- Methadone
- Mushrooms
- Nicotine
- Volatile substance
- Semeron (fictitious drug)

Each of these drug variables can take 6 different values:

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

Data Pre-Processing

Encoding columns into numeric data & One Hot Encoding

```
for column in col_drogue:
    le = LabelEncoder()
    df[column] = le.fit_transform(df[column])

for column in col_démographie:
    le = LabelEncoder()
    df[column] = le.fit_transform(df[column])

for column in col_personnalité:
    le = LabelEncoder()
    df[column] = le.fit_transform(df[column])
```

```
oh_data = pd.get_dummies(data_regulier, columns = ['Age', 'Education'])

oh_data.drop(['Age_2.59171'], axis=1, inplace = True)

oh_data.rename(columns = {'Age_-0.95197': 'Age: 18-24',
                          'Age_-0.07854': 'Age: 25-34',
                          'Age_0.49788': 'Age: 35-44',
                          'Age_1.09449': 'Age: 45-54',
                          'Age_1.82213': 'Age: 55-64',
                          'Education_-2.43591': 'Décrochage avant 16 ans',
                          'Education_-1.7379': 'Décrochage à 16 ans',
                          'Education_-1.43719': 'Décrochage à 17 ans',
                          'Education_-1.22751': 'Décrochage à 18 ans',
                          'Education_-0.61113': 'Ecole supérieure ou Université',
                          'Education_-0.05921': 'Certificat professionnel',
                          'Education_0.45468': 'Diplômé universitaire',
                          'Education_1.16365': 'Diplômé de master',
                          'Education_1.98437': 'Diplômé de doctorat'
                          }, inplace = True)
```

Dropping irrelevant feature columns

Dropping rows where people answered they took the fictitious drug (Semeron) to identify overclaimers and exclude their other answers

Dropping fictitious drug column for the rest of the analysis

Data Pre-Processing for classification

Binary Classification Problem for each drug :

```
def tester(f):  
    if ((f==6) or (f==5) or (f==4) or (f==3) or (f==2) or (f==1)):  
        f = 1  
    elif (f==0):  
        f = 0  
    return f
```

```
def regulier(f):  
    if ((f==6) or (f==5)):  
        f = 1  
    elif ((f==0) or (f==1) or (f==2) or (f==3) or (f==4)):  
        f = 0  
    return f
```

```
data_test=data.copy()  
for col in col_droque:  
    data_test[col]=data_test[col].map(tester)
```

Tested the drug at least once (value 1) :

- CL1 : Used over a Decade
- CL2 : Used in the Last Decade
- CL3 : Used in the Last Year
- CL4 : Used in the Last Month
- CL5 : Used in the Last Week
- CL6 : Used in the Last Day

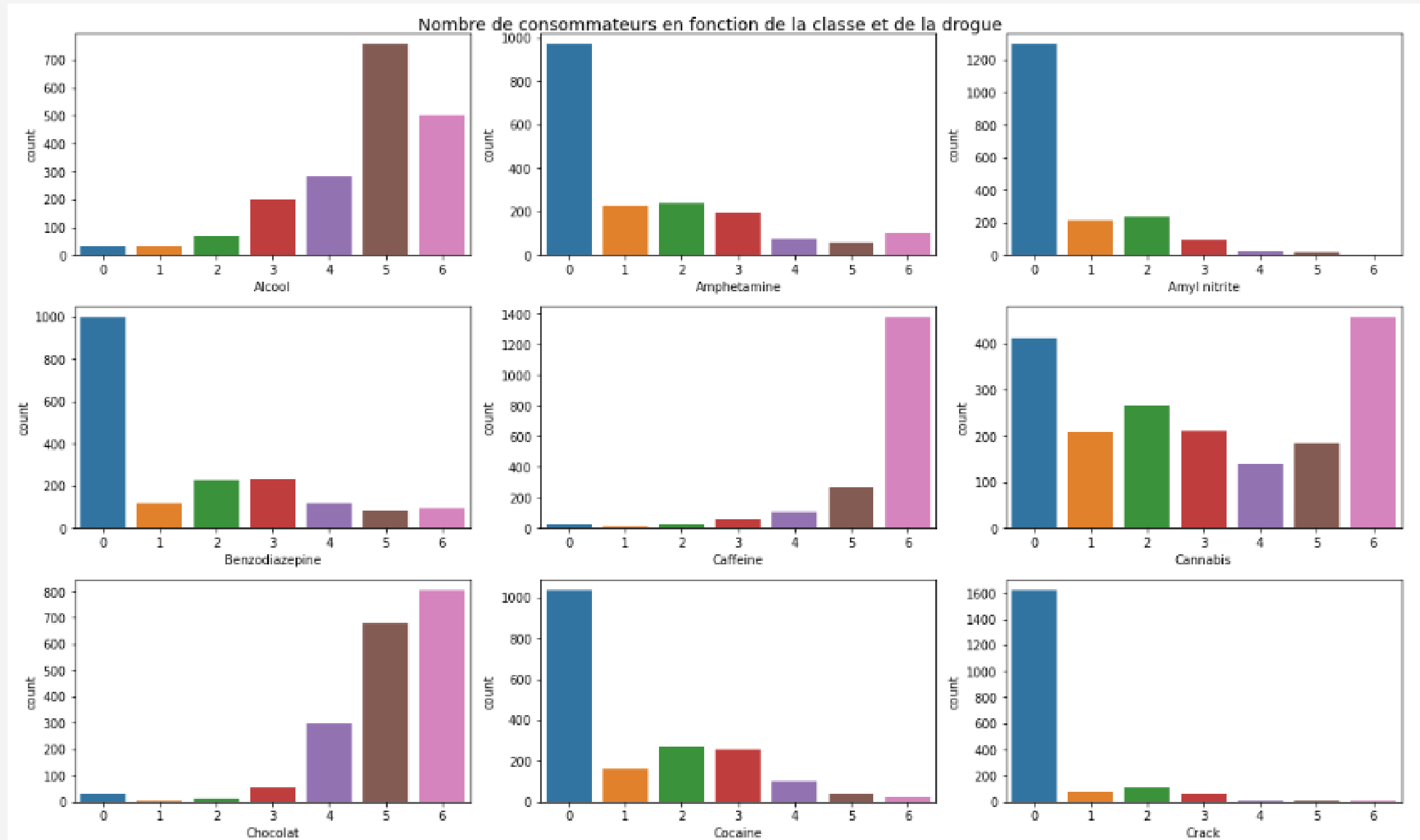
Never tested the drug (value 0) :

- CL0 : Never Used



Data Visualizations

Visualizations of the number of users by category for each drug

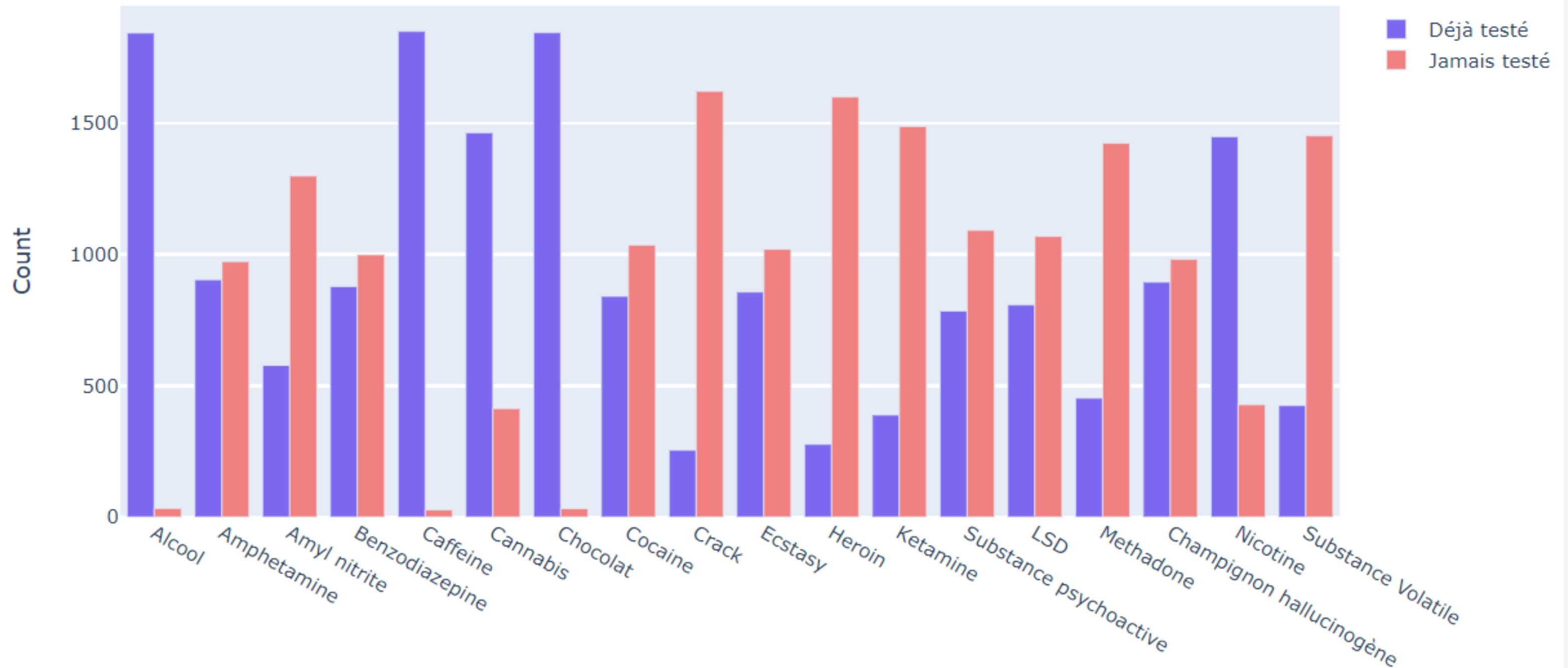




Data Visualizations

Visualizations of the number of people who tested or not each drug

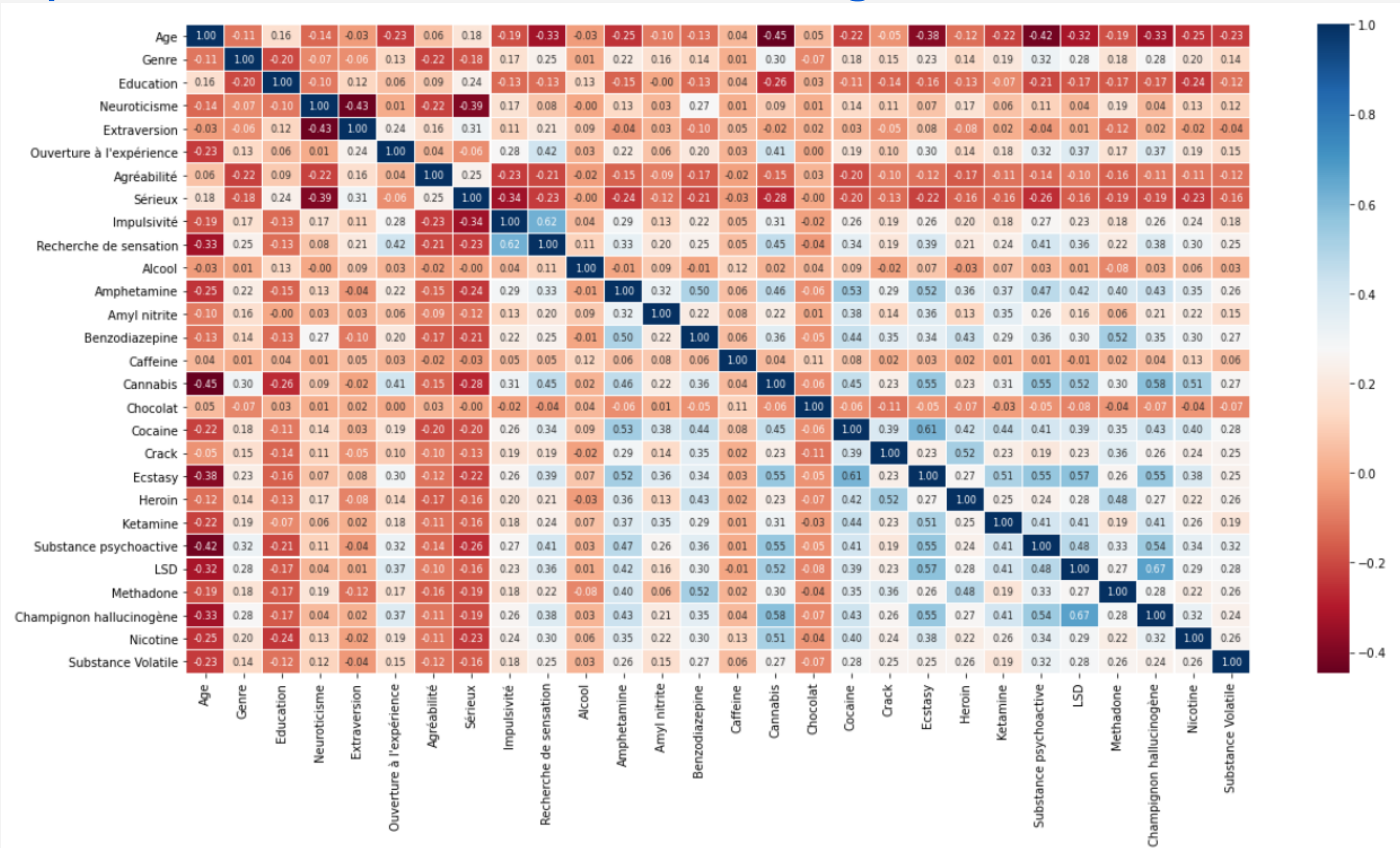
Nombre de personnes ayant déjà testé ou non chaque drogue





Data Visualizations

Heatmap : Correlations between each feature and drug





Data Modeling

Predicting whether an individual has ever tested a drug or not

A few algorithms tested to analyze Cannabis, Ecstasy, Mushrooms and LSD consumption

▲ **Logistic regression**

▲ **KNN**

▲ **Decision Tree**

▲ **Support Vector Machine**

▲ **Random Forest**

Data Modeling

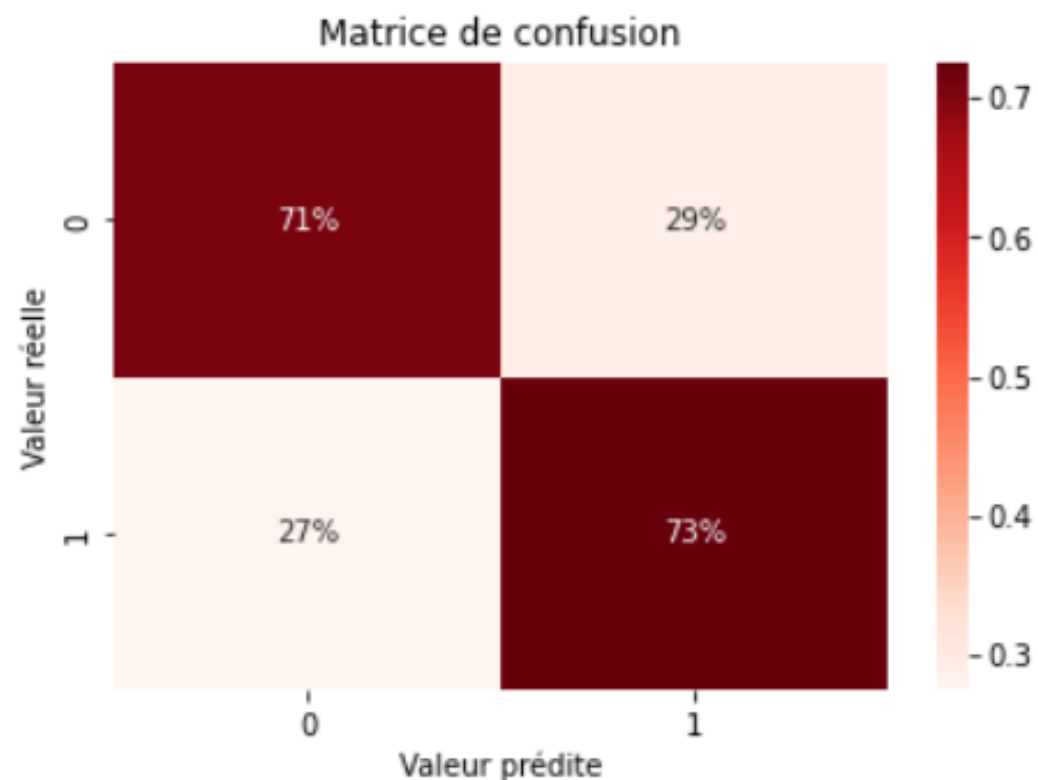
Predicting whether an individual has ever tested ecstasy or not

Grid Search to find the best parameters

```
Best params:
{'criterion': 'entropy', 'max_depth': 7, 'max_features': 'auto', 'n_estimators': 500}
Train f1 score: 0.696
Test f1 score: 0.704
```

	precision	recall	f1-score	support
0	0.75	0.71	0.73	302
1	0.68	0.73	0.70	262
accuracy			0.72	564
macro avg	0.72	0.72	0.72	564
weighted avg	0.72	0.72	0.72	564

Confusion Matrix

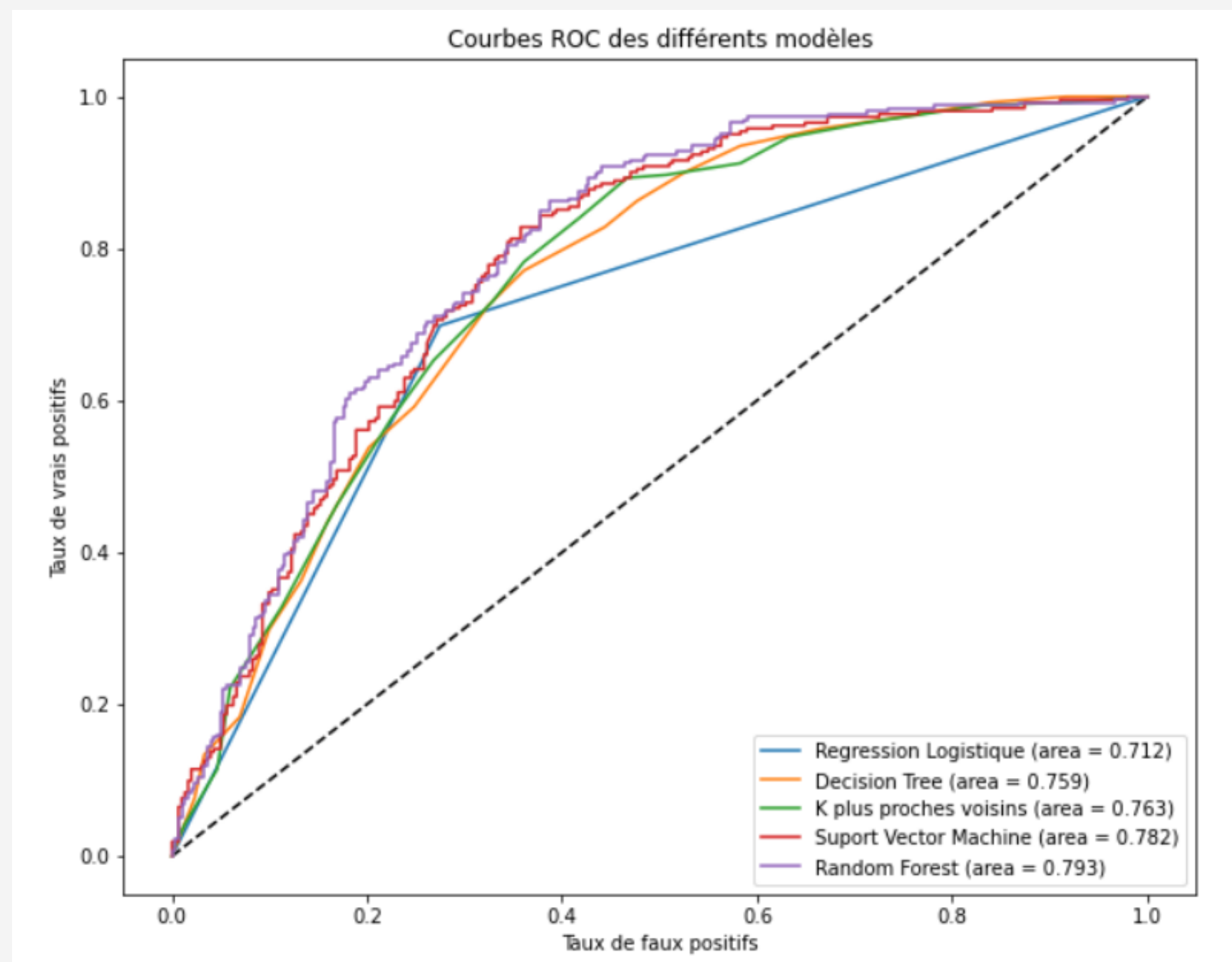




Data Modeling

Predicting whether an individual has ever tested ecstasy or not

Comparison of the models with ROC curves



F1 Scores of the models

	f1-score
Random Forest	0.703704
Suport Vector Machine	0.697936
K plus proches voisins	0.692029
Regression Logistique	0.691871
Decision Tree	0.690909

Data Modeling

Same steps for other target variables

Results for Cannabis :

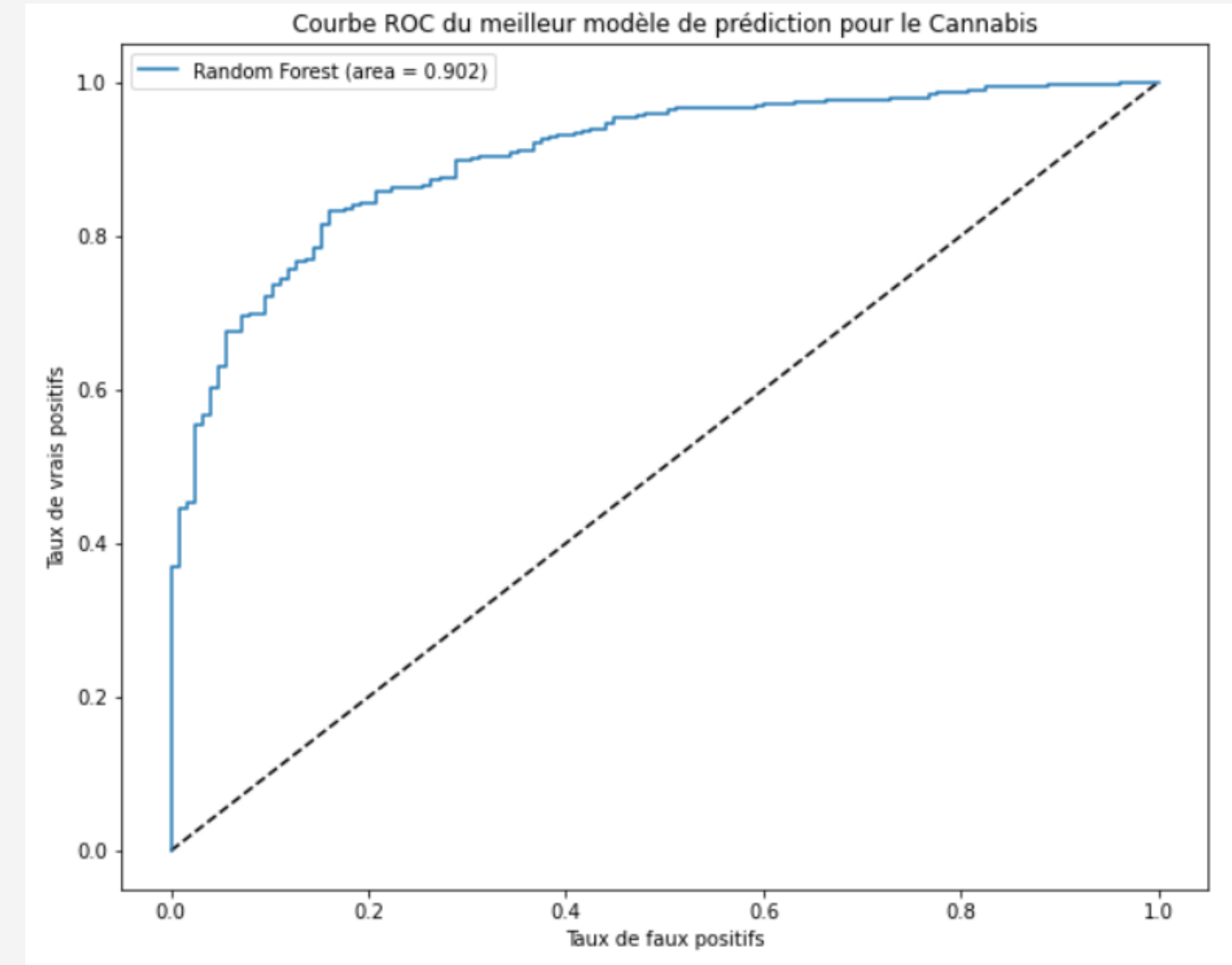
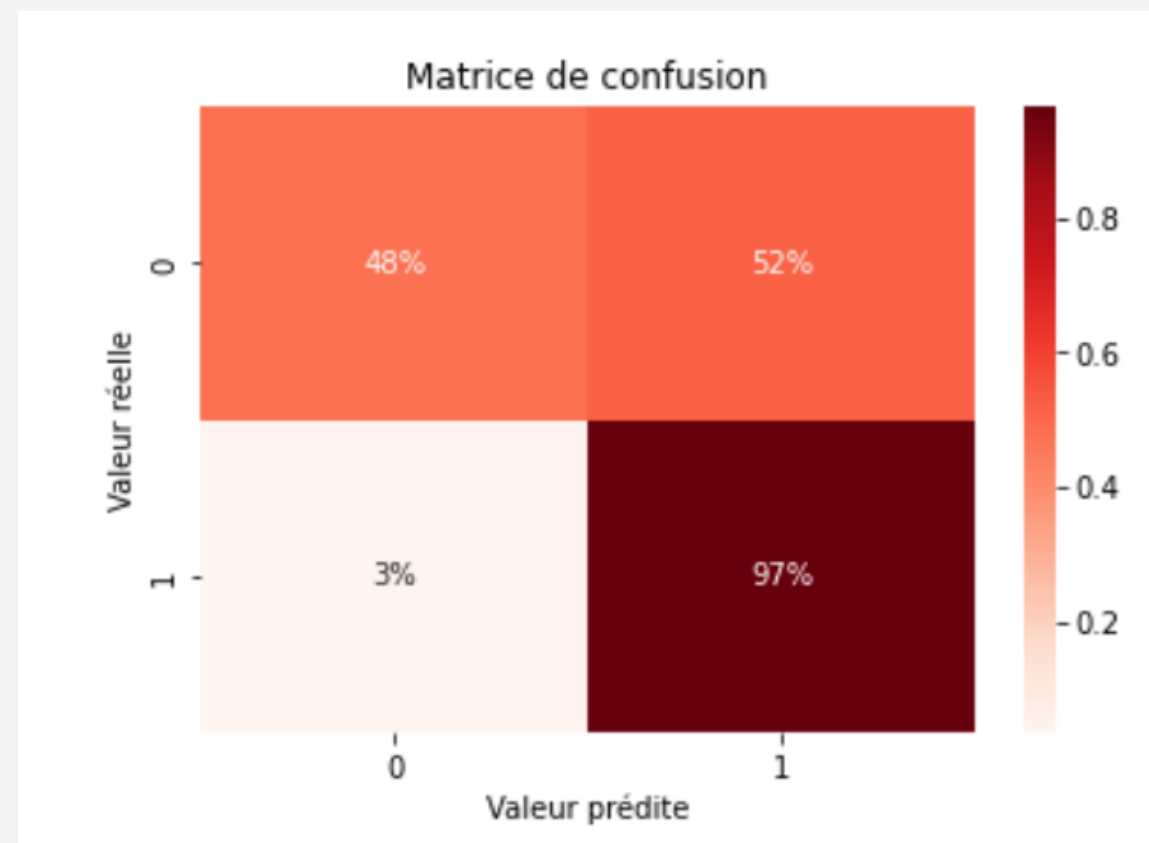
Best params:

```
{'criterion': 'entropy', 'max_depth': 5, 'max_features': 'auto', 'n_estimators': 500}
```

Train f1 score: 0.911

Test f1 score: 0.914

	precision	recall	f1-score	support
0	0.80	0.48	0.60	125
1	0.87	0.97	0.91	439
accuracy			0.86	564
macro avg	0.83	0.72	0.76	564
weighted avg	0.85	0.86	0.84	564



Thank you !

Cyprien NICOLAY - Timothé VITAL - Anna ZENOU
DIA 5