

A complex network graph is visible in the background, composed of numerous small white dots (nodes) connected by thin white lines (edges). Some nodes are highlighted in light blue. In the upper left, there's a large, dense cluster of nodes. To the right, several individual nodes are connected by lines to form small triangles. A few larger triangles are also present. The overall effect is one of data connectivity and complexity.

# Python for data analysis - Projet

---

Drug Consumption (quantified) dataset



Présentation du  
dataset

01

Papier de Recherche  
et problématique  
d'étude

02

Exploration du  
dataset et des  
variables

03

## TABLE OF CONTENTS

04

Machine Learning

05

API

# 01

## Présentation du dataset



# 1. Présentation du dataset

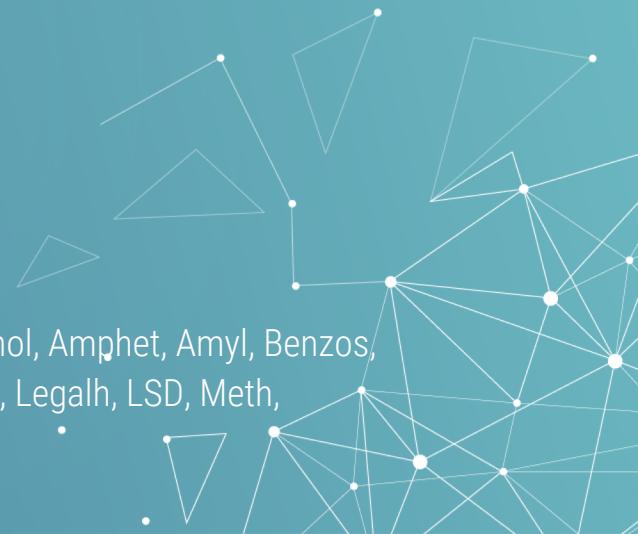
Source : <https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29>

- Ce dataset est le résultat d'un sondage concernant la consommation de drogue.
- 1885 personnes ont participé et répondu à des questions concernant la consommation de 18 drogues différentes
- On retrouve pour chaque personne des scores correspondant au « Five Factor Model of personality », cinq domaines qui définissent une personnalité
- Dans le dataset de départ , les indicateurs sont codés par des nombres
- Chaque nombre correspond à une valeur définie dans la documentation du dataset



# Variables du dataset

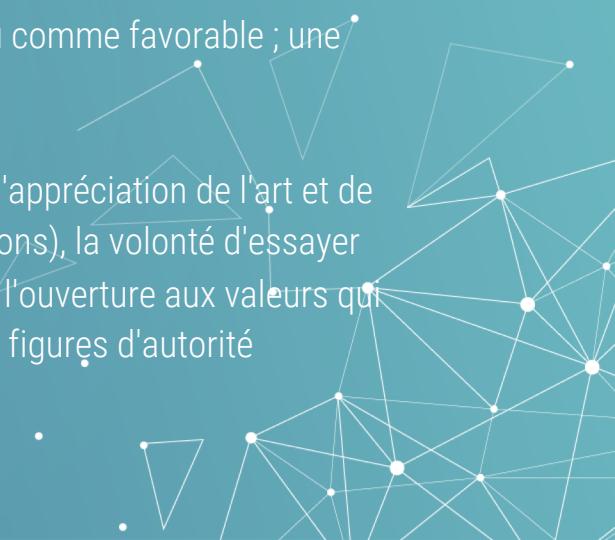
- ID
- Age : correspond à une catégorie d'âge
- Gender : genre du participant
- Education : catégories d'études réalisées
- Country : pays d'origine
- Ethnicity
- Nscore : NEO-FFI-R Neuroticism score
- Escore : NEO-FFI-R Extraversion score
- Oscore : NEO-FFI-R Openness to experience score
- Ascore : NEO-FFI-R Agreeableness score
- Cscore : NEO-FFI-R Conscientiousness score
- Impulsive : Impulsiveness score measured by BIS-11
- SS :Sensation seeking score measured by ImpSS
- Consommation et fréquence de consommation des 18 drogues : Alcohol, Amphet, Amyl, Benzos, Caffeine, Cannabis, Chocolate, Coke, Crack , Ecstasy, Heroin, Ketamyn, Legalh, LSD, Meth, Mushrooms, Nicotine, Semer, VSA



# Personnalité et Scores (1)

Les scores du dataset correspondent à des domaines définissant une personnalité.

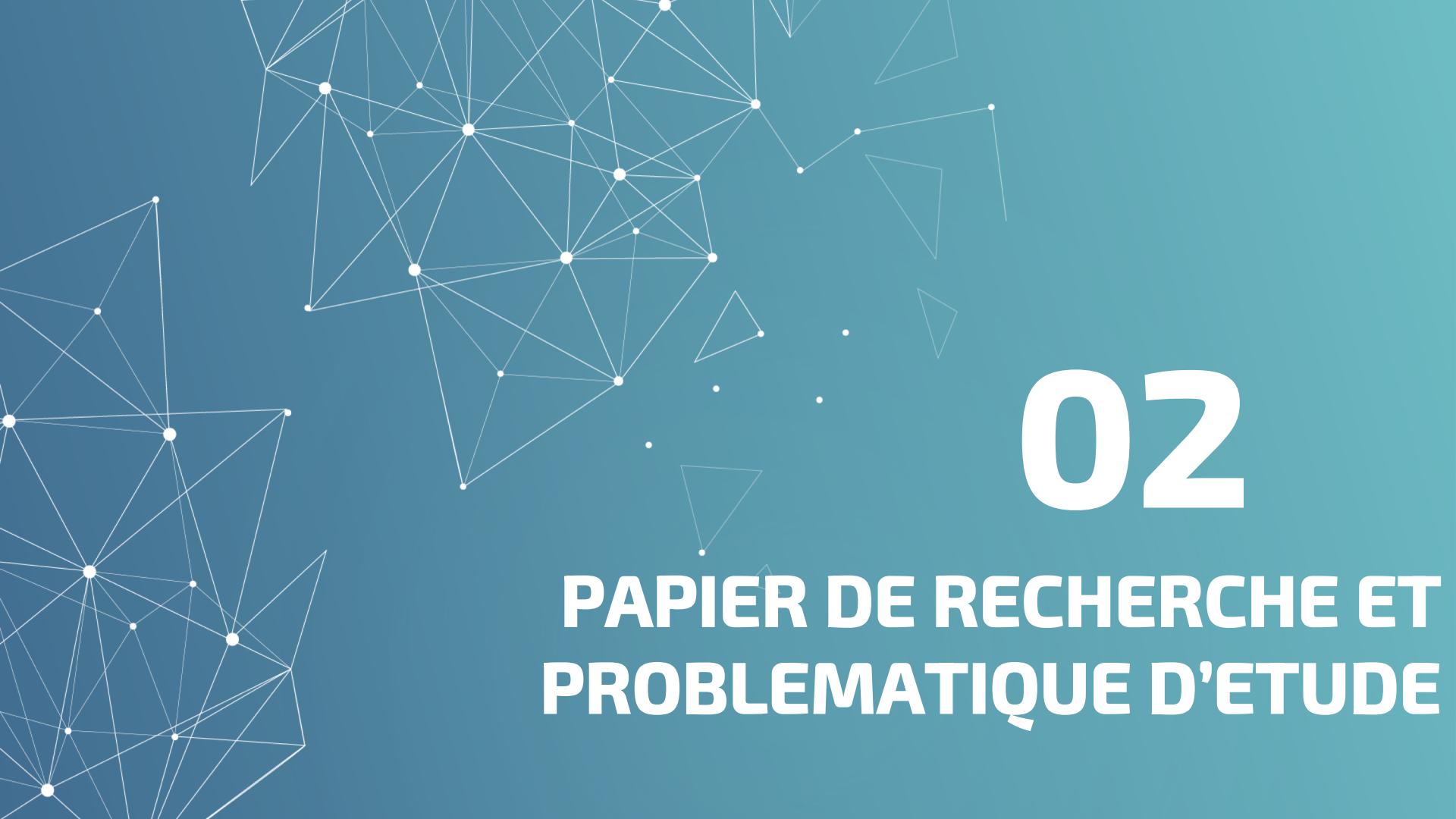
- Nscore (Neuroticism) : Les individus possédant un haut degré de neuroticisme peuvent faire l'expérience d'émotions telles que l'anxiété, la colère, la culpabilité et la déprime
- Escore (Extraversion) : Un score élevé en extraversion indique une forte réactivité aux *stimuli* agréables, traduite par une tendance à percevoir, construire et ressentir la réalité et les événements comme stimulants et agréables. L'environnement est perçu comme favorable ; une source de plaisir qu'il faut aller cueillir.
- Oscore (openness to experience) : correspond à l'ouverture aux rêveries, l'appréciation de l'art et de la beauté (esthétique), l'ouverture aux sentiments (réceptivité aux émotions), la volonté d'essayer des activités différentes et nouvelles, la curiosité intellectuelle ainsi que l'ouverture aux valeurs qui permet de remettre en question ses propres valeurs ainsi que celles des figures d'autorité



# Personnalité et Scores (2)

- Ascore (agréabilité) : un score élevé sur cette dimension ont tendance à croire que la plupart des gens sont honnêtes, décents et digne de confiance.
- Cscore : Les personnes conscientieuses sont généralement efficaces et organisées par opposition à d'autres qui privilégieraient le travail vite fait et désorganisé.
- Ils sont complétés par un score d'impulsivité et de recherche de sensation.



A complex network graph is visible in the background, composed of numerous white dots (nodes) connected by thin white lines (edges). The nodes are of varying sizes, creating a sense of depth and connectivity. Some nodes are highlighted with a larger size, suggesting they are central or have higher degrees of connectivity.

# 02

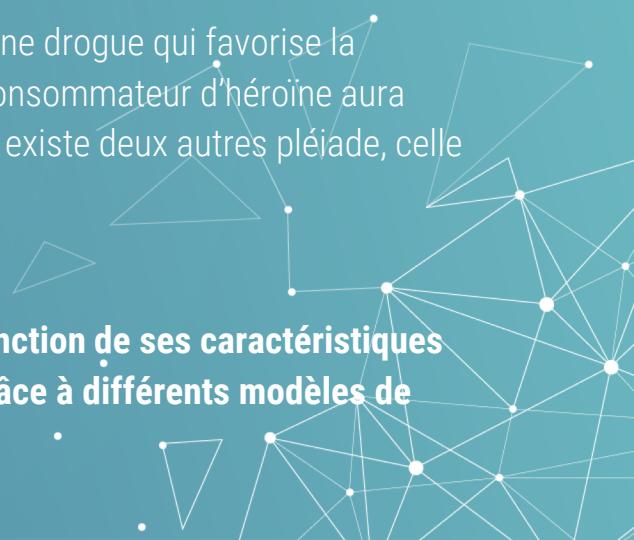
## PAPIER DE RECHERCHE ET PROBLEMATIQUE D'ETUDE

# Article de recherche

Papier de recherche : <https://arxiv.org/pdf/1506.06297v2.pdf>

Des chercheurs ont donc travaillé sur ce dataset. Dans leur article, il présentent leur démarche et leurs résultats.

- Le but de cette étude est de prévenir le risque de consommation de drogue chez des individus.
- L'étude montre qu'on a une corrélation entre les scores concernant la personnalité et la probabilité de consommer certaines drogues.
- Elle met aussi en évidence une corrélation entre la consommation d'une drogue qui favorise la consommation d'une autre. Par exemple la pléiade de l'heroïne : un consommateur d'héroïne aura tendance à aussi consommer du crack, de la cocaine et de la meth (il existe deux autres pléiade, celle de l'ecstasy et celle des benzodiazepines).
- L'objectif est donc le suivant :
  - **Prévoir si un individu consommera ou pas une drogue en fonction de ses caractéristiques psychologiques mais aussi de ses autres consommations grâce à différents modèles de Machine Learning.**



# 03

## EXPLORATION DU DATASET ET DES VARIABLES

# Exploration du dataset

Comment le dataset a-t-il été exploré ?

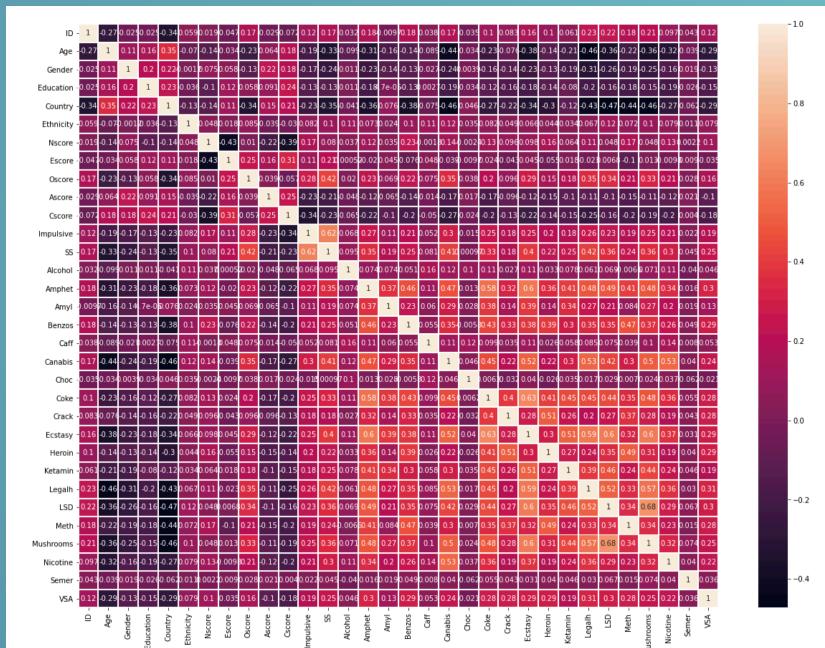
- Etape 1 : exploration des données personnelles des individus. On regarde dans quelle catégorie d'age, d'études, de genre et de pays ils se situe pour mieux connaître la population étudiée
- Etape 2 : Fréquence de consommation de chaque drogue. Afin de comprendre comment, à quelle fréquence et dans quelles proportions sont consommées chaque drogue, on étudie ce paramètre.
- Etape 3 : étude plus poussée de certaines drogues (Cafféine, Ecstasy, Cannabis) pour trouver des facteurs qui influent sur la consommation. On a écarté les facteurs personnels (age, éducation) qui n'ont pas donné beaucoup d'informations. On remarque qu'il y a une différence dans la répartition des scores (Nscore, Escore ...) entre les consommateurs et non consommateurs d'une drogue.
- Etape 4 : on fait une matrice de corrélation pour avoir pour chaque éléments les facteurs qui ont une importance. C'est cette matrice qui va servir de base au choix des paramètres pour le machine learning.

# Résultats de la matrice de corrélation

- On remarque d'abord que les scores psychologiques ont une influence et un poids différent selon les drogues consommés. On peut donc dire que c'est un facteur intéressant à utiliser pour l'apprentissage.
  - On peut aussi confirmer via cette matrice l'existence des pleiade (on retrouve les mêmes que celles mentionnées par l'article) :

- Héroïne avec crack meth et cocaine
  - Ecstasy avec amphetamines, cannabis, cocaine, ketamine, LSD, magic mushrooms, legal highs
  - Benzos avec methadone, amphetamines, cocaine

Utiliser ces pléiades dans l'apprentissage sera aussi intéressant



# 04

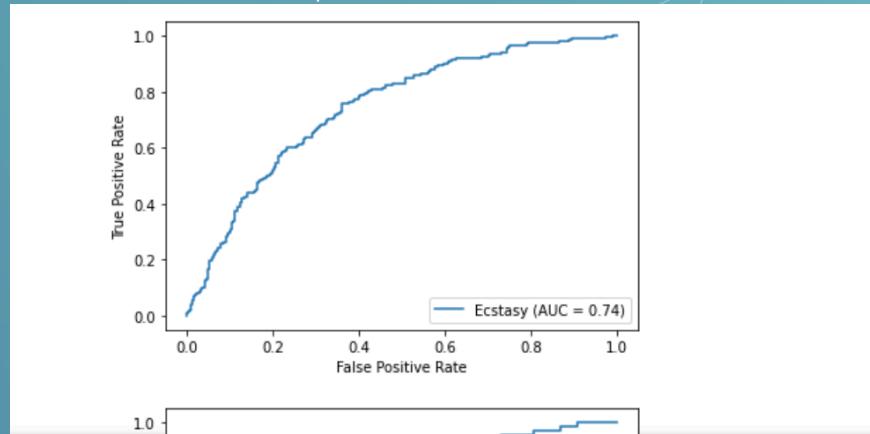
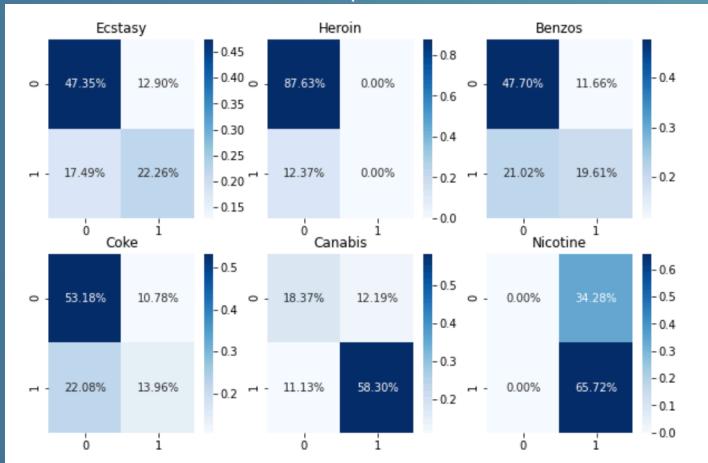
# MACHINE LEARNING

---



# Prédiction en fonction des scores (1)

- Suite au lien que l'on a fait précédemment entre les scores et la probabilité de consommer certaines drogue on veut créer des modèles de prédiction en fonction de ces paramètres.
- On va créer des modèles grâce à la librairie sklearn et les tester sur un petit échantillon de drogues. Le but est de voir si les prédictions sont efficaces et si les algorithmes marchent aussi bien pour des drogues différentes. Voici les modèles implémentés : SVC, Naive Bayes et k nearest neighbors. Pour chaque modèle on affiche les matrices de corrélation ainsi que les courbes ROC afin de conclure sur l'efficacité de chaque modèle. Voici un exemple des résultats obtenus pour un des modèles :



# Prédiction en fonction des scores (2)

- Quelle conclusion tire-t-on de cette première partie de machine learning ?

Il est donc possible est assez efficace de créer des modèles avec notre dataset. Ces modèles permettent, à partir des paramètres scores psychologiques d'un individu, de prévoir s'il est probable qu'il consomme telles drogues. Ici on a testé avec 6 drogues différentes et les résultats sont plutôt bons.

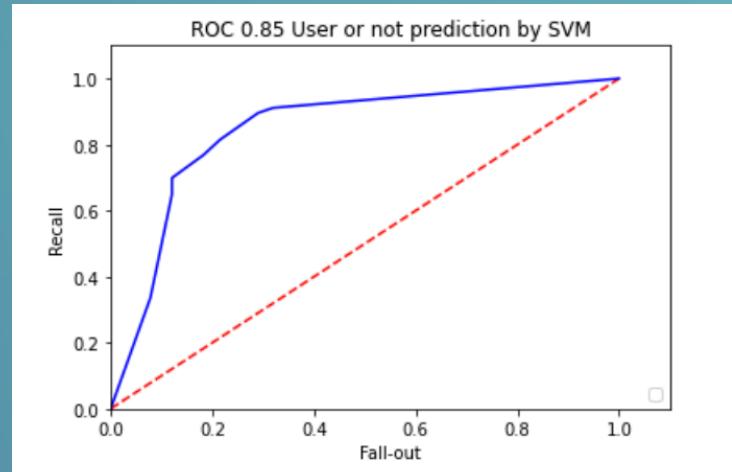
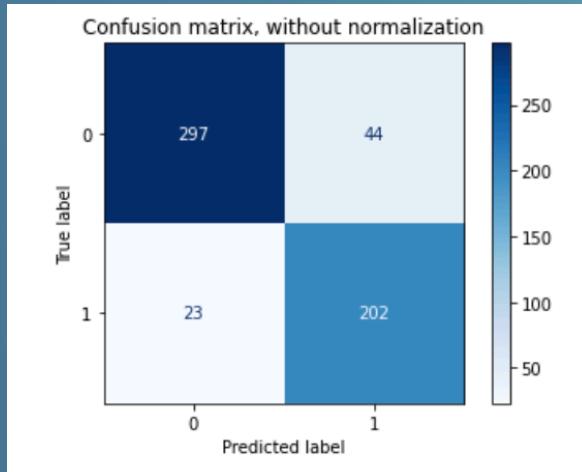
On remarque qu'un même modèle n'est pas forcément le meilleur pour chaque drogue. Il faudrait donc définir quelle méthode donne les meilleurs résultats sur chaque drogue et ainsi créer le modèle le plus optimal pour chaque si on voulait les utiliser dans des domaines médicaux par exemple.

Les chercheurs ayant travaillé sur ce dataset mettent aussi en évidence ceci dans leurs travaux.



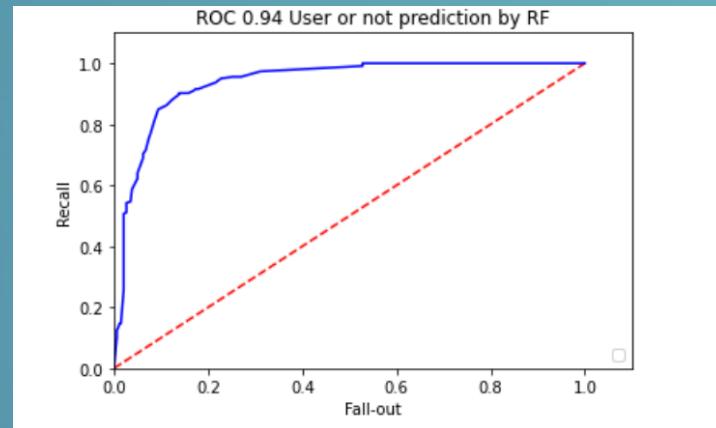
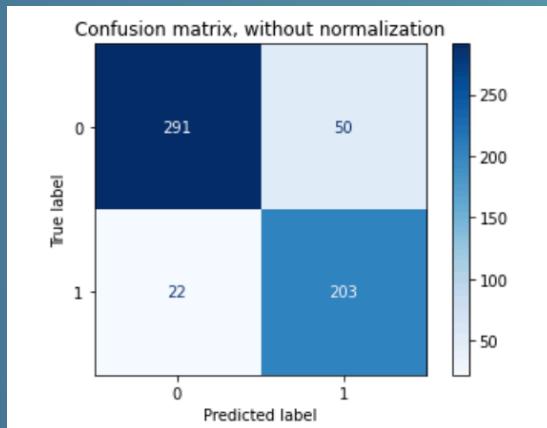
# Prédiction en fonction des autres drogues (1)

- Ici on utilise le concept des pléiades afin de prédire si un individu va consommer telle drogue ou non
- On se concentre sur l'ecstasy : on sait donc que les autres drogues révélatrices sont 'Amphet', 'Canabis', 'Coke', 'Ketamin', 'LSD', 'Mushrooms' et 'Legalh'. Ce sont les paramètres utilisés pour la suite
- 1<sup>er</sup> algorithme : SVM gridsearch



# Prédiction en fonction des autres drogues (2)

- 2<sup>e</sup> algorithme : random forest grid search



- Ces deux algorithmes permettent de faire une assez bonne prédiction. Cela est dû au fait qu'on choisi des paramètres qui ont une corrélation importante (on le savait grâce à la matrice de corrélation et à la théorie). Ces algorithmes permettent donc eux aussi d'effectuer des prédictions intéressantes.

# 05

## API



# API Flask

- On crée une API Flask. Via celle-ci, il est possible de faire des prédictions via le model svc pour l'Ecstasy en entrant les différents scores (nscore, escore ...).
- On exporte le modèle depuis le code source python grâce à `joblib.dump(svc_ec, filename)`.
- L'API run en exécutant le script app.py
- En tant qu'utilisateur, il faut remplir la page à l'adresse <http://0.0.0.0:3333> avec les scores :

The screenshot shows a web browser window titled "IRIS PREDICTOR". The page contains six input fields labeled "nscore", "escore", "oscore", "ascore", "cscore", and "impulsive", each with a spin-up/spin-down control. Below these inputs is a "Predict" button. At the top right of the browser window, there are links to "Apple", "Google", "YouTube", and "Gmail".

