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

5 demographic features :

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

7 personality features :

- Neuroticism
- Extraversion
- Opennes 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

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

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
- Dropping irrelevant feature columns
- Dropping rows where people answered they took the ficticious drug (Semeron) to identify overclaimers and exclude their other answers
- Dropping ficticious drug column for the rest of the analysis

Data Pre-Processing

for classification

Binary Classification Problem for each drug :

Non Regular User (value 0) :

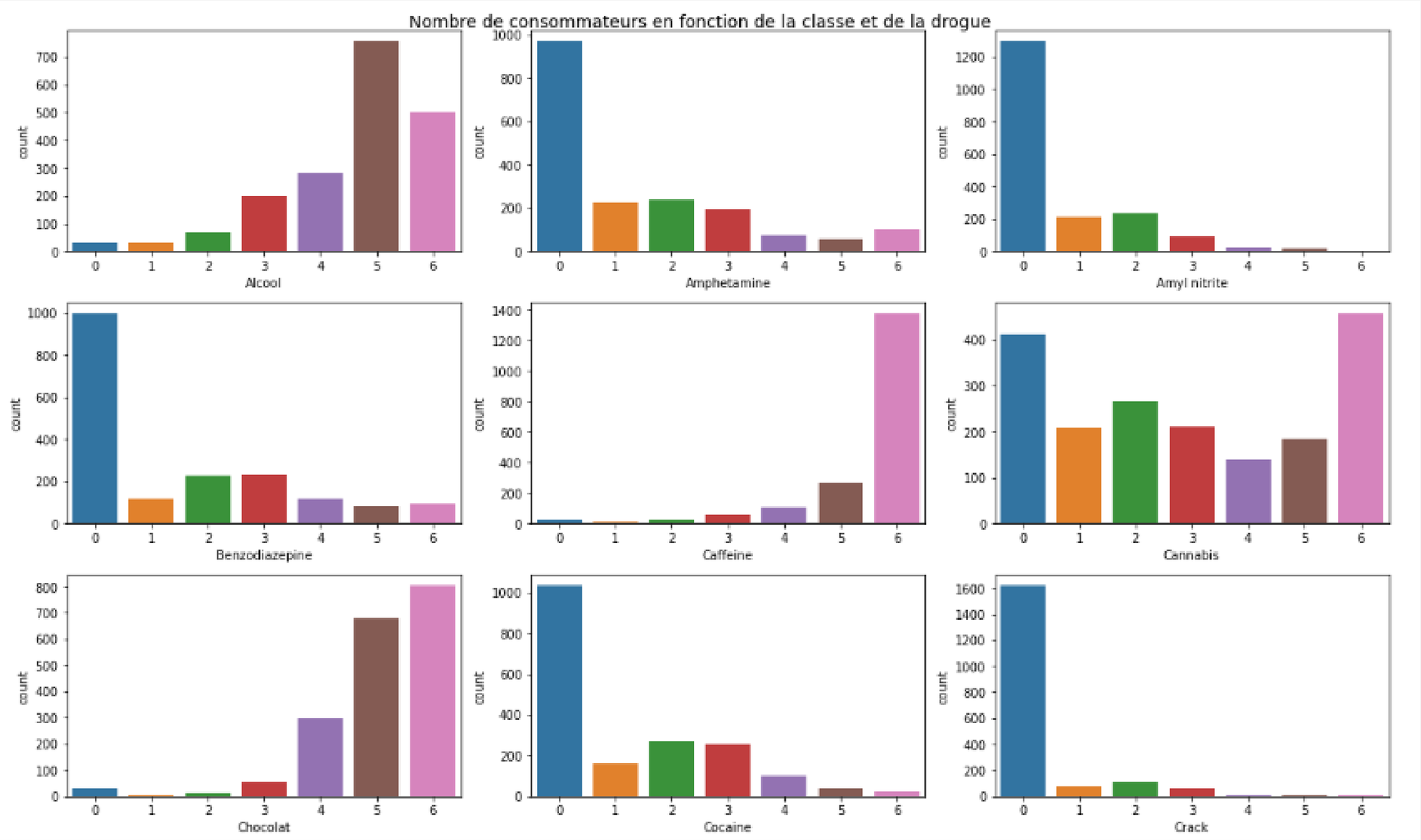
- CL0 : Never Used
- CL1 : Used over a Decade
- CL2 : Used in the Last Decade
- CL3 : Used in the Last Year
- CL4 : Used in the Last Month

Regular User (value 1) :

- CL5 : Used in the Last Week
- CL6 : Used in the Last Day

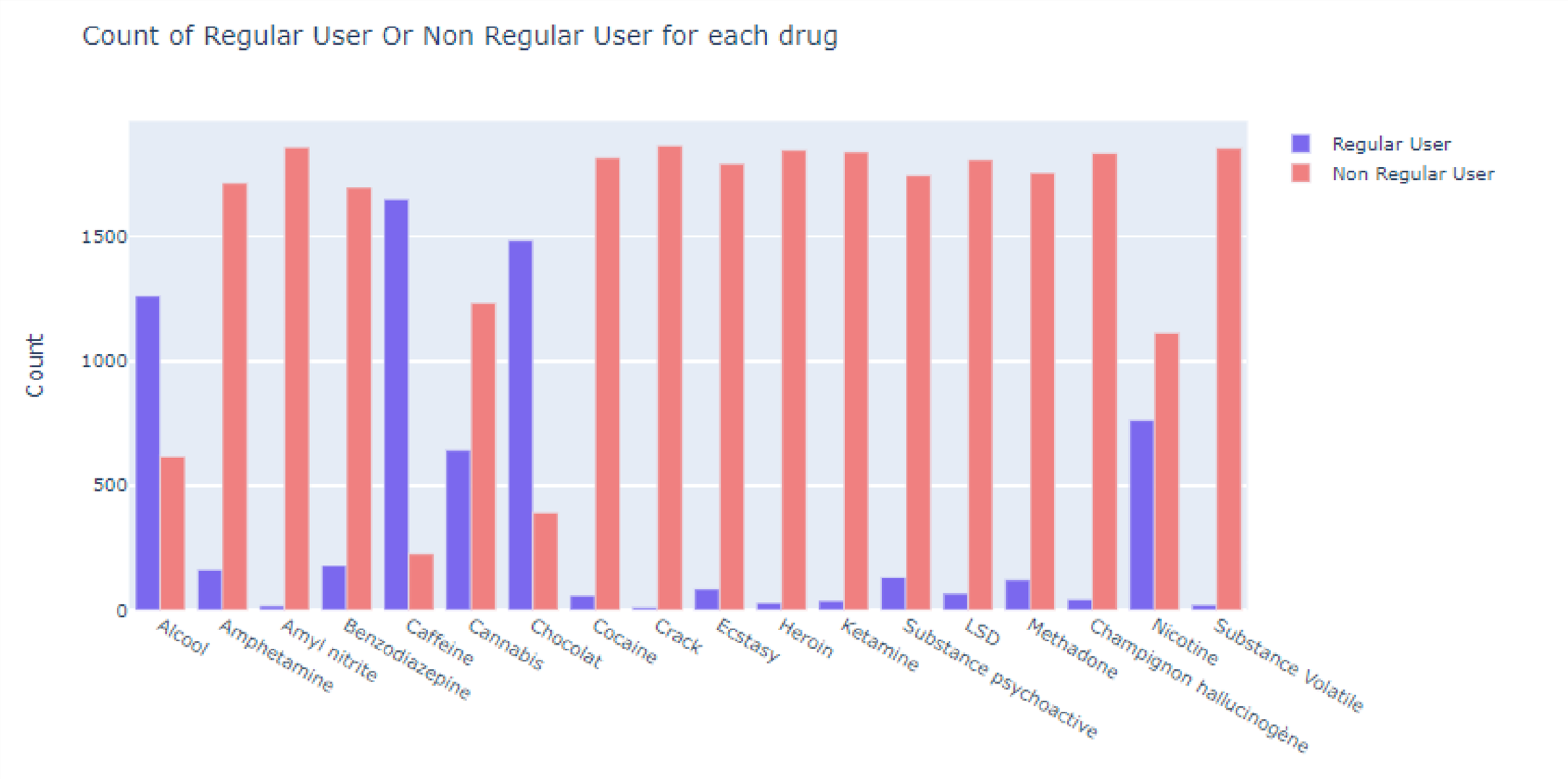
Data Visualizations

Visualizations of the number of users by category for each drug



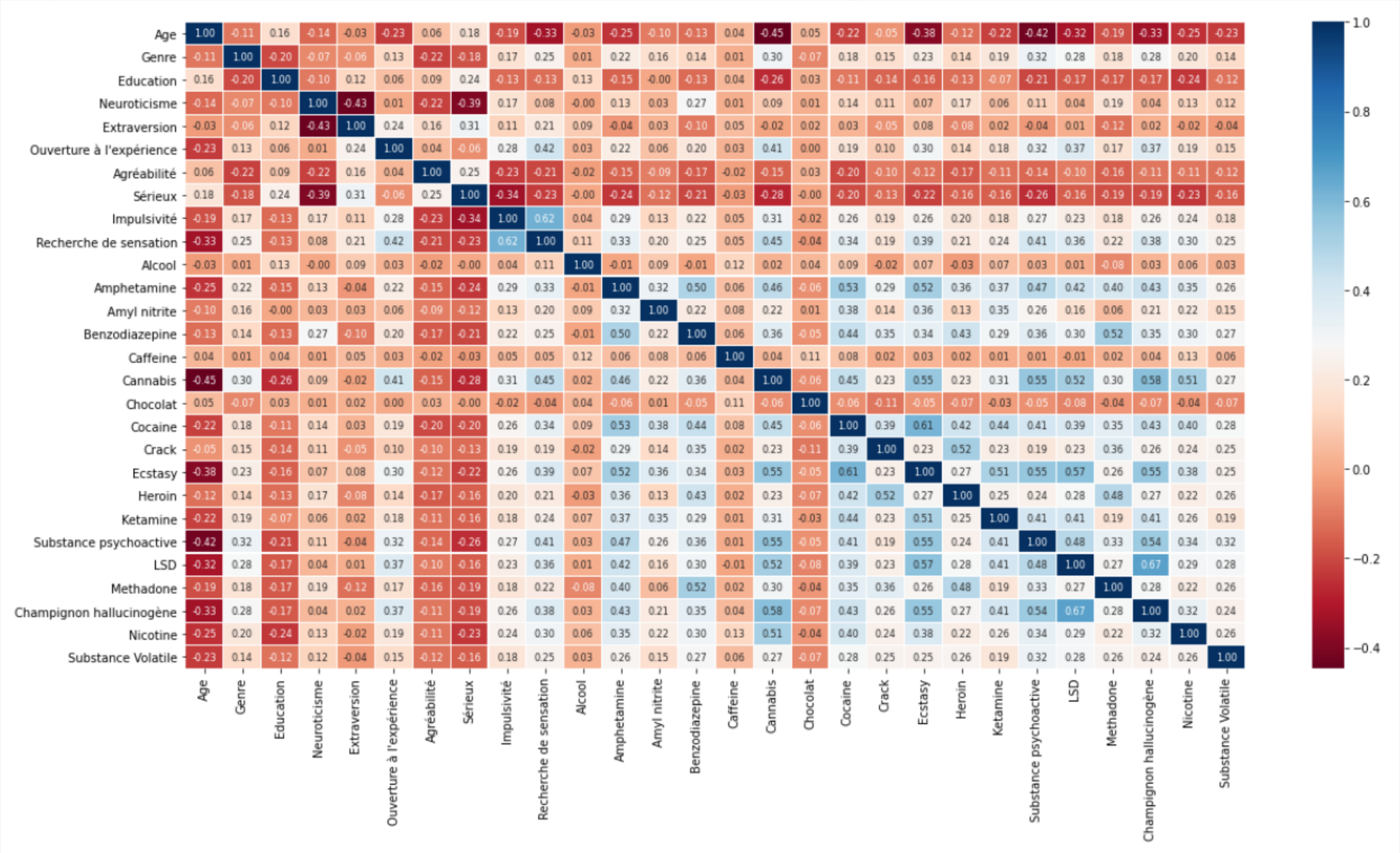
Data Visualizations

Visualizations of the number of regular and non regular user for each drug



Data Visualizations

Heatmap : Correlations between each feature and drug



Data Modeling

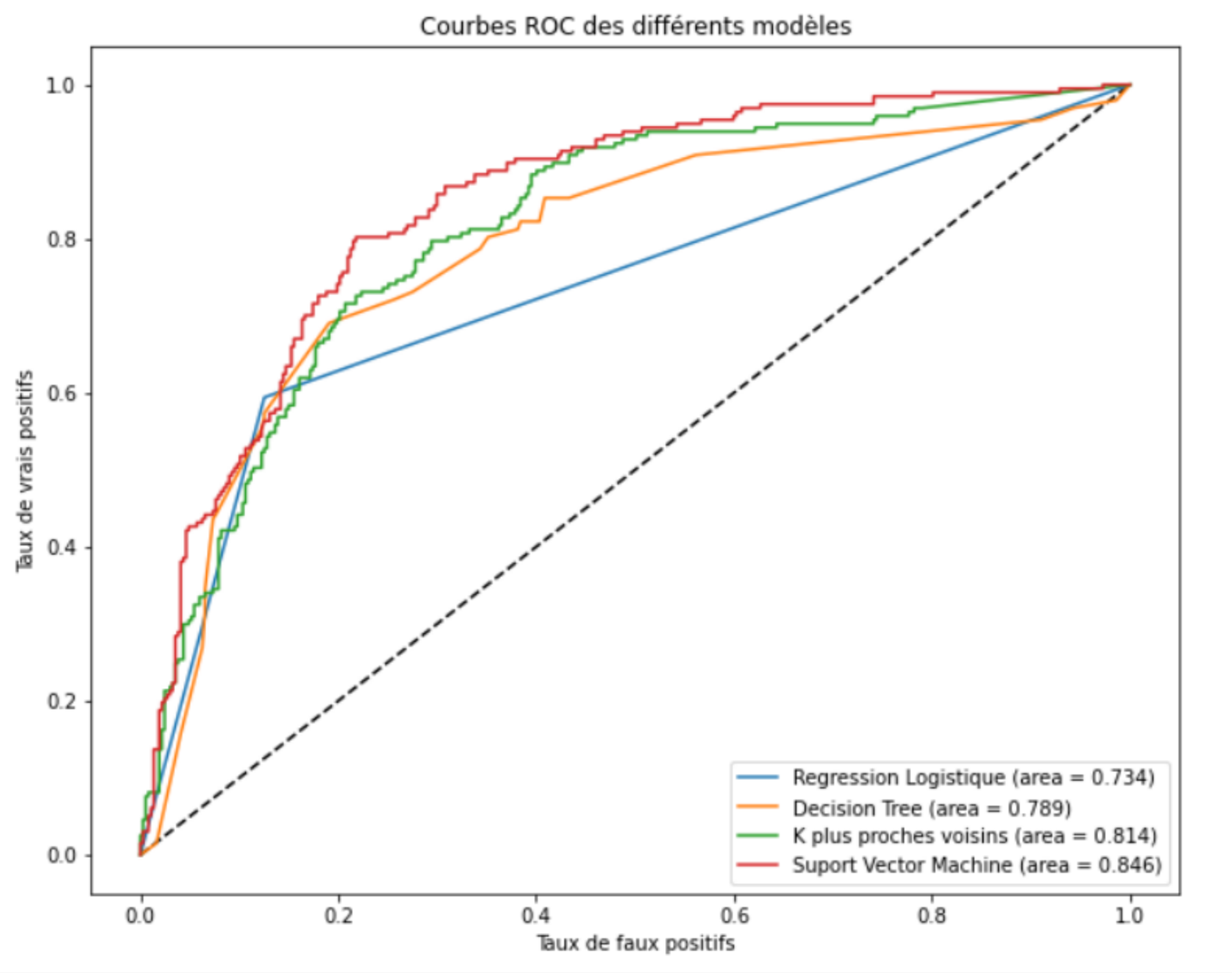
Predicting whether an individual is a regular or non regular user

A few algorithms tested to analyze Cannabis consumption

- **Logistic regression**
- **KNN**
- **Decision Tree**
- **Support Vector Machine**

Data Modeling

Predicting whether an individual is a regular or non regular user



Comparison of the models with ROC curve

API : DJANGO

Thank you !

Cyprien NICOLAY - Timothé VITAL - Anna ZENOU
DIA 5