DRUG CONSUMPTION

Valentin ALAKILETOA-PINAULT
Safia ALIOUCHE
Elies BENMALEK

INTRODUCTION

- This project aims to study the "Drug Consumption" dataset and to set up a machine learning model to predict whether or not an individual consumes drugs.

- To do this we will :
  - Reading data
  - Data cleaning
  - Data visualization
  - Setting up the models
  - Comparison of models

DATA-PREPROCESSING

READ FILES AND RENAME COLUMNS

- 1885 rows
- 32 columns

	0	1	2	3	4	5	6	7	8	9	...
0	1	0.49788	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	...
1	2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	...
2	3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	...
3	4	-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	...
4	5	0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	...

https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified

	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	Ascore	Cscore
0	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	-0.00665
1	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	-0.14277
2	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	-1.01450
3	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	0.58489
4	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	1.30612

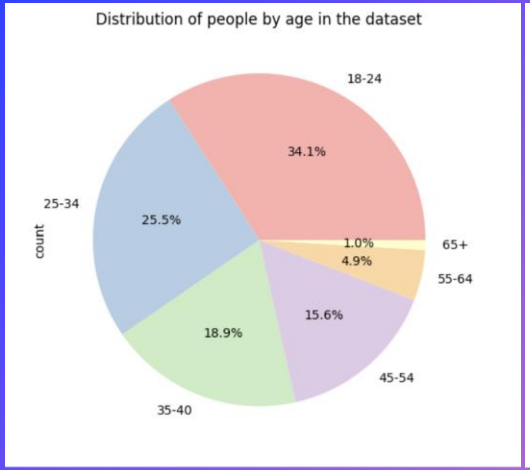
DATA CLEANING

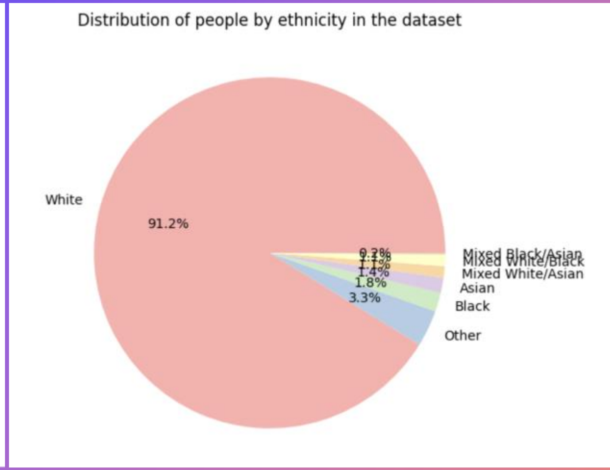
	ID	Age	Gender	Education	Country	Ethnicity
0	1	35-40	F	Professional certificate/ diploma	UK	Mixed White/Asian
1	2	25-34	M	Doctorate degree	UK	White
2	3	35-40	M	Professional certificate/ diploma	UK	White
3	4	18-24	F	Masters degree	UK	White
4	5	35-40	F	Doctorate degree	UK	White

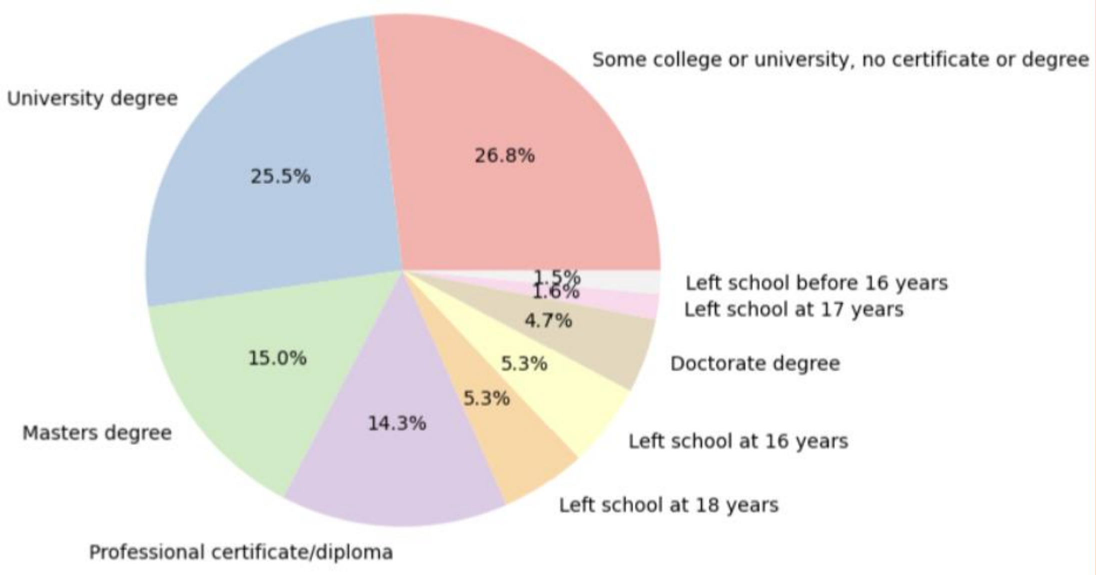
Ecstasy	Heroin	Ketamin	Legalh	LSD	Meth	Mushrooms	Nicotine	Semer	VSA
Never	Never	Never	Never	Never	Never	Never	Last Decade	Never	Never
Last Month	Never	Last Decade	Never	Last Decade	Last Year	Never	Last Month	Never	Never
Never	Never	Never	Never	Never	Never	Decade Ago	Never	Never	Never
Never	Never	Last Decade	Never	Never	Never	Never	Last Decade	Never	Never
Decade Ago	Never	Never	Decade Ago	Never	Never	Last Decade	Last Decade	Never	Never

Using the dataset archives, we changed all the values by their meaning.

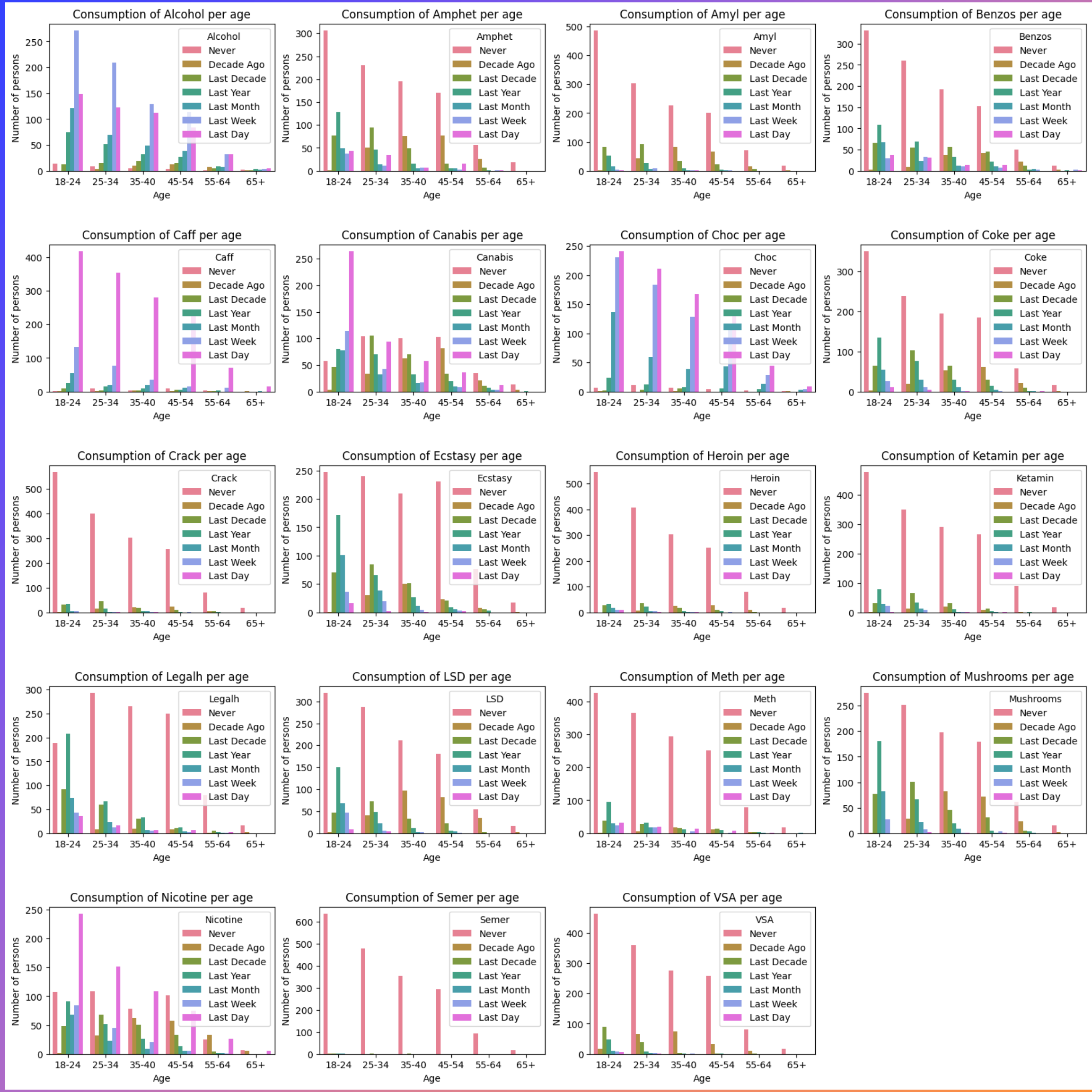
DATA VISUALIZATION





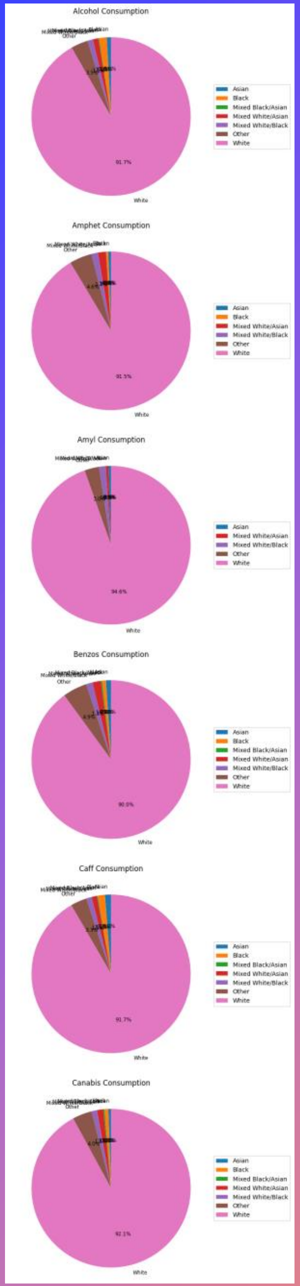


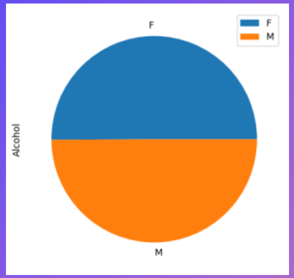
- First of all, we wanted to have a visual on the distribution of individuals according to their characteristics in the data set.
- As we can see there are certain characteristics that are much more represented than others.
- For example, there are almost only white people in the dataset.

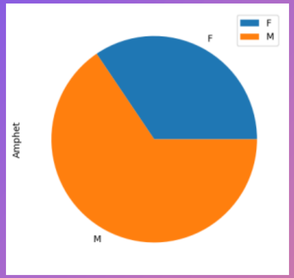


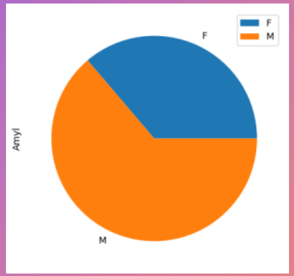
- We then wanted to know what and how the individuals in the dataset consume according to the group of people they belong to.

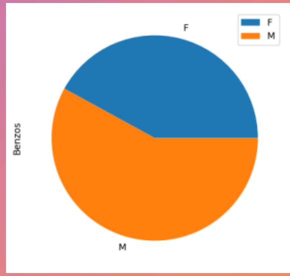
- We can see for example that almost everyone has tasted alcohol and chocolate, regardless of the age. A lot of people tried Canabis too, but less people have tried Coke.





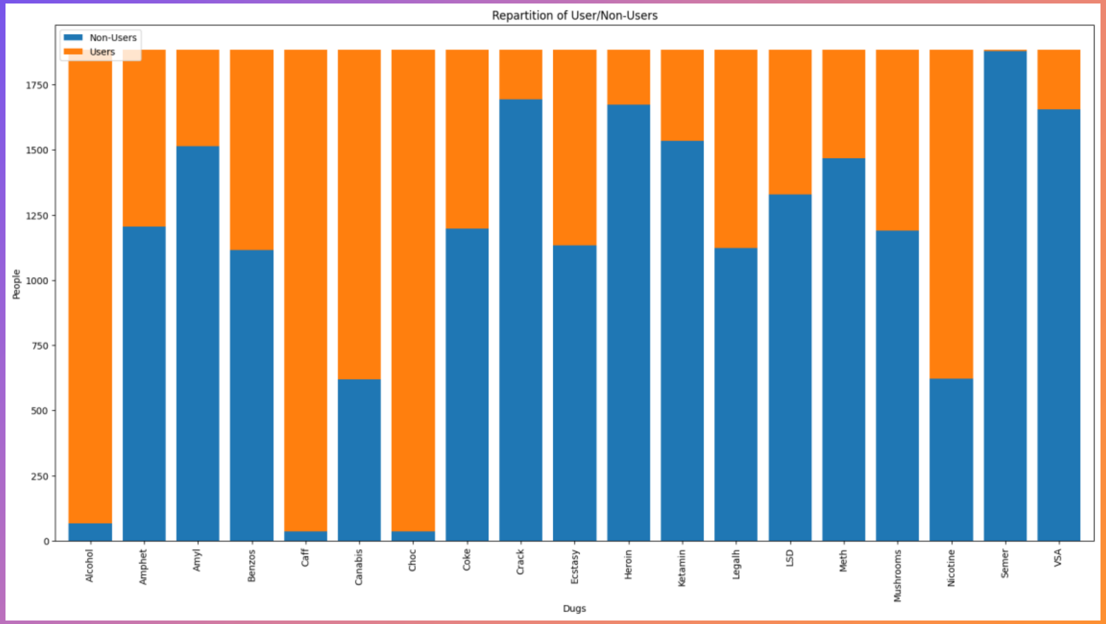






- Because the distribution between each category in the dataset is disproportionate, these plots are unusable. As seen previously, the fact that there are only white people in the dataset distorts the results by giving the impression that being white implies consumption.
- Only the Gender plot is usable, since its distribution is evenly split in the dataset

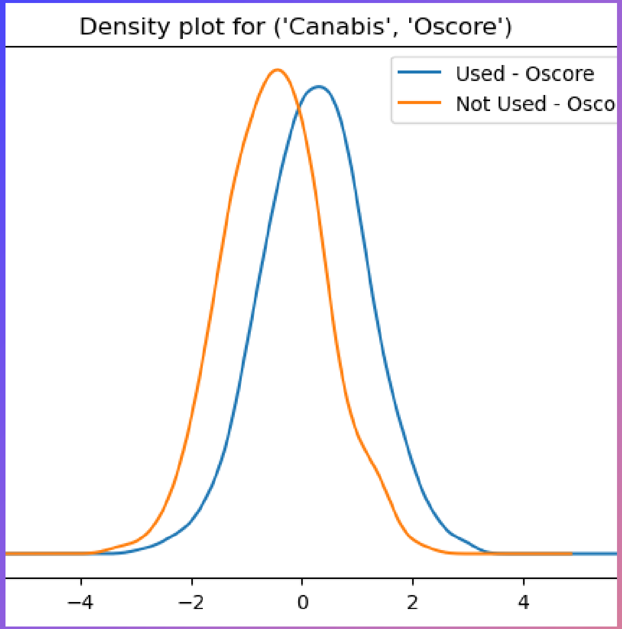
Ecstasy	Heroin	Ketamin	Legalh	LSD	Meth	Mushrooms	Nicotine	Semer	VSA
0	0	0	0	0	0	0	1	0	0
1	0	1	0	1	1	0	1	0	0
0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	1	0	0
0	0	0	0	0	0	1	1	0	0
...	...	...	...	...	...	...	...	...	...
0	0	0	1	1	0	0	0	0	1
1	0	0	1	1	1	1	1	0	0
1	0	1	0	1	0	1	1	0	0
1	0	0	1	1	0	1	1	0	0
1	0	0	1	1	0	1	1	0	1

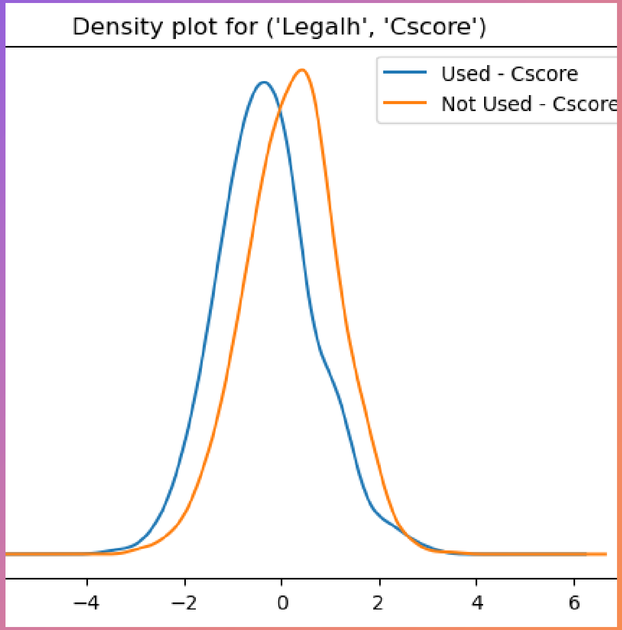


- We made two parts in the dataset: consumers (1) and non-consumers (0)

- We considered that an individual was not a user if he had never used the drug or not in the last 10 years.

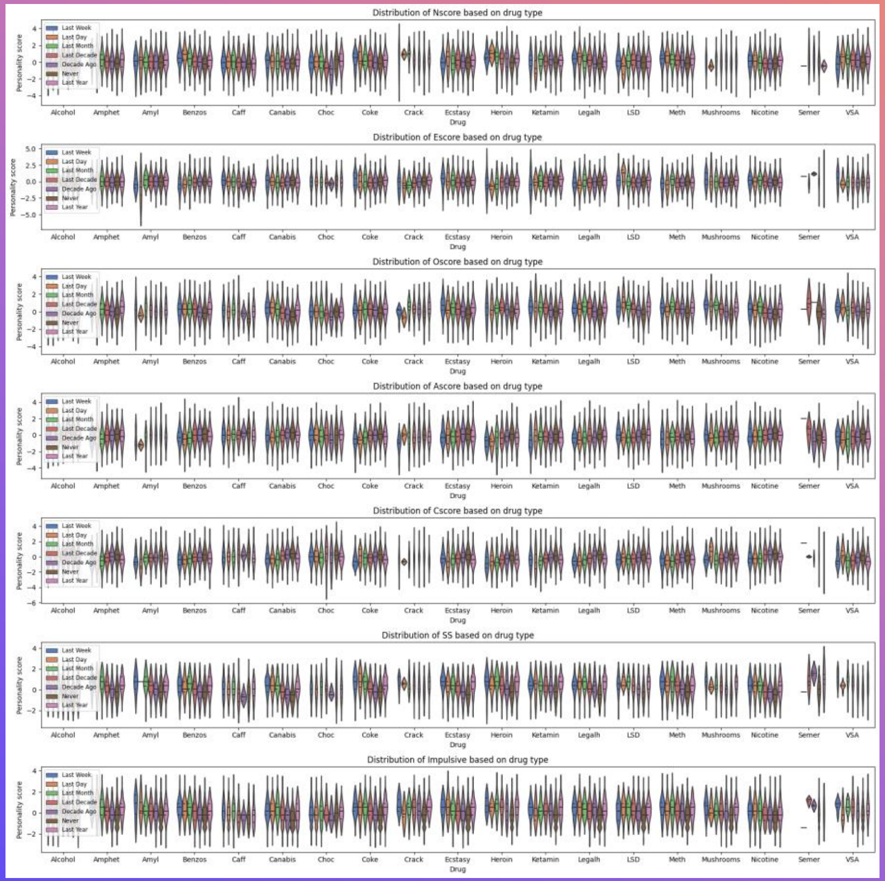
- Which brings us to this graph of the distribution of users and non-users according to each drug.

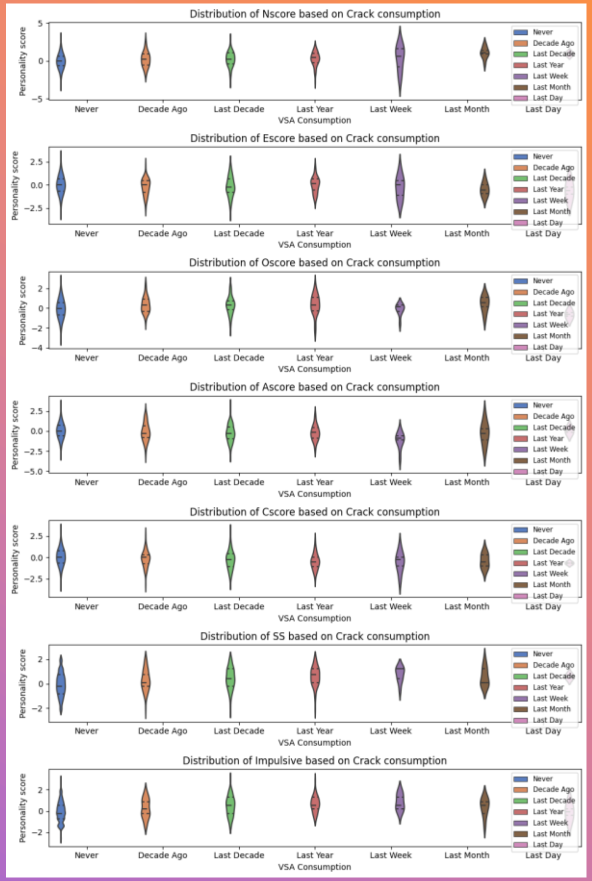




- Finally, we plotted the density of the scores to see if they were distributed well enough in the dataset so that we could use them to predict whether an individual is a consumer or not.

- We can clearly see that the distribution of scores for Users and non-Users is not the same. Thus, the scores make it possible to obtain information to classify an individual in one of the two categories.





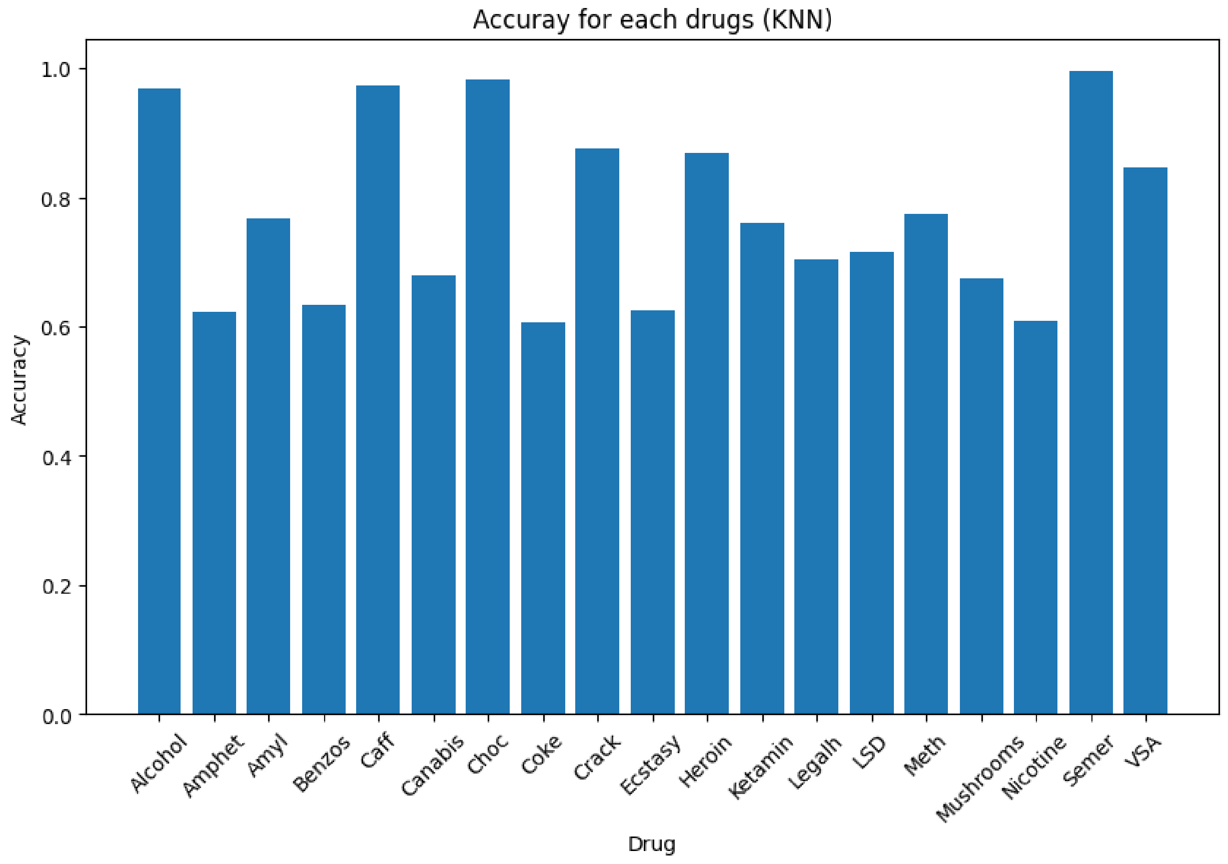
These plots provides us many observations, for example : Individuals who regularly consume crack tend to be particularly neurotic. Crack has an impact on openness in individuals who regularly use them. People who take crack experience a sense of agreeableness that diminishes after a week.

MODELISATION

KNN

WHY : KNN can be applied effectively if our dataset is not very large and if we have a suitable selection of relevant features.

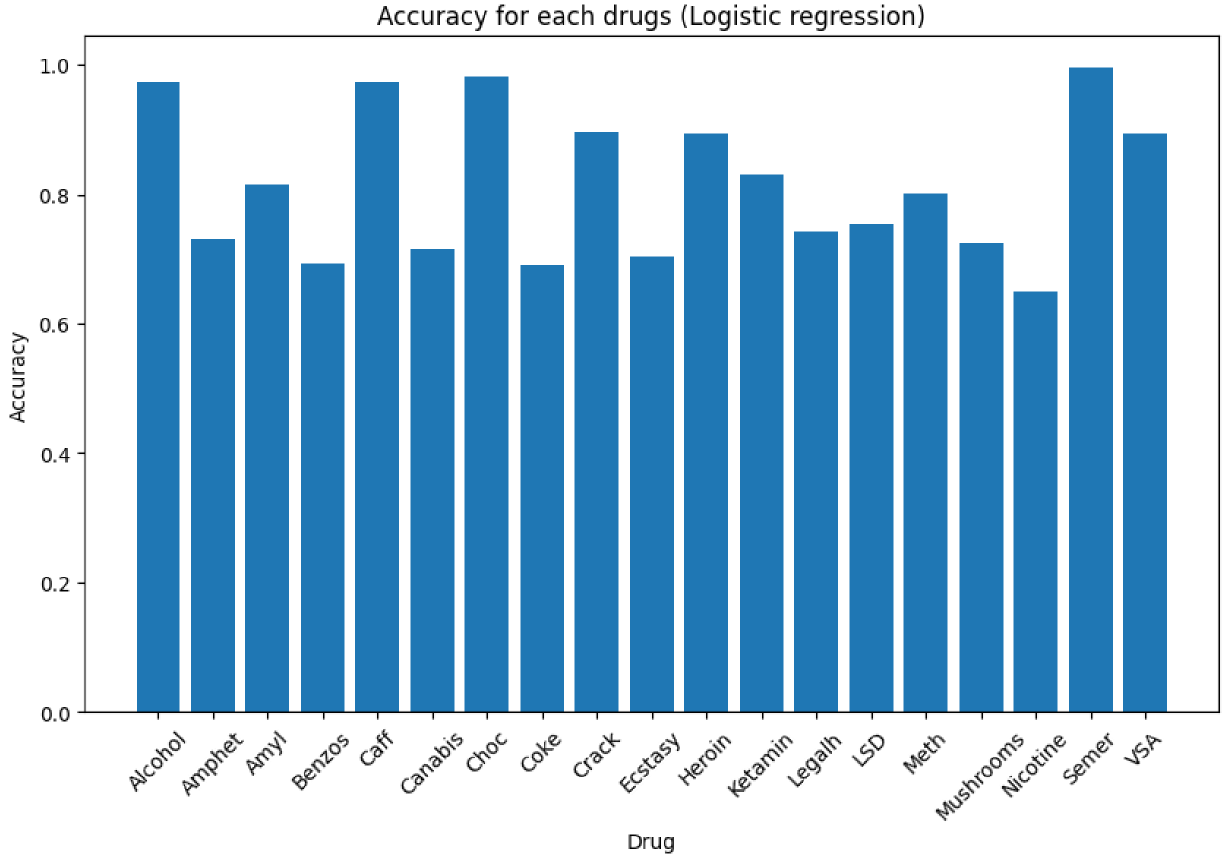
```
Alcohol : 0.9681697612732095
Amphet : 0.623342175066313
Amyl : 0.76657824933687
Benzos : 0.6339522546419099
Caff : 0.9734748010610079
Canabis : 0.6790450928381963
Choc : 0.9814323607427056
Coke : 0.6074270557029178
Crack : 0.8753315649867374
Ecstasy : 0.6259946949602122
Heroin : 0.870026525198939
Ketamin : 0.7612732095490716
Legalh : 0.7029177718832891
LSD : 0.7161803713527851
Meth : 0.7745358090185677
Mushrooms : 0.6737400530503979
Nicotine : 0.610079575596817
Semer : 0.9946949602122016
VSA : 0.8461538461538461
```



Logistic Regression

WHY : Logistic regression is suitable for predicting binary variables (0 or 1) based on continuous and categorical characteristics.

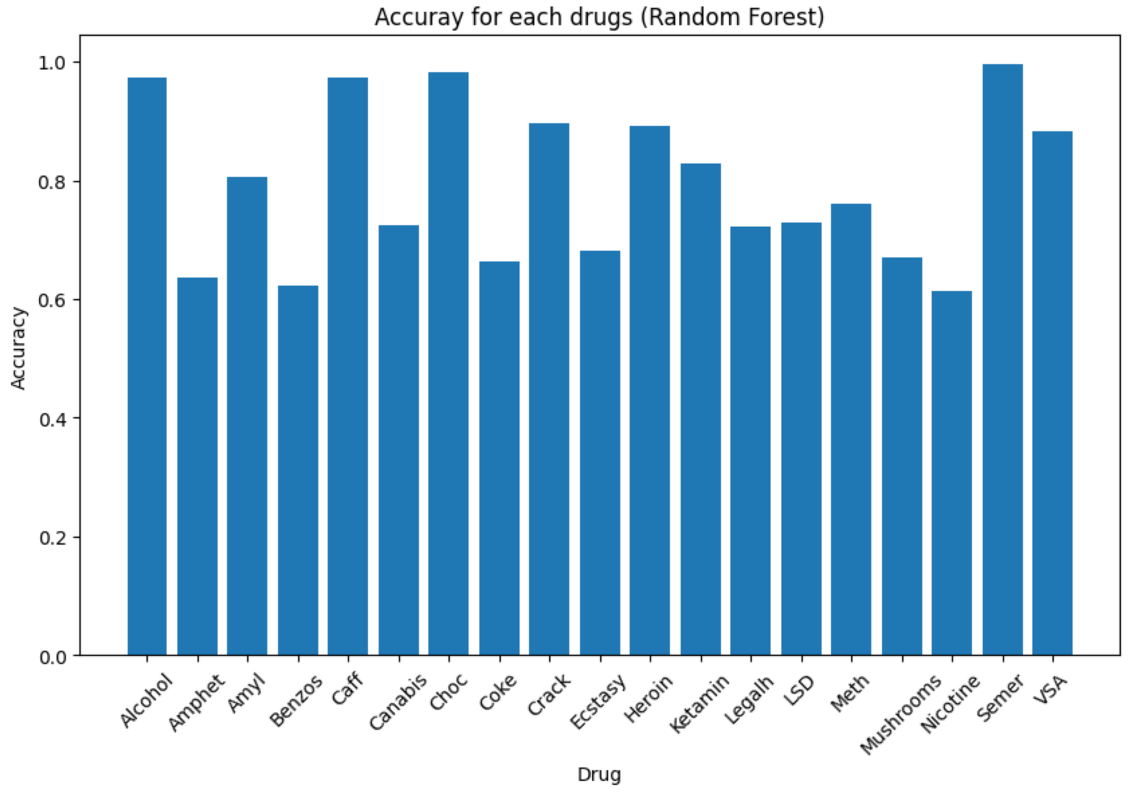
```
Alcohol : 0.9734748010610079
Amphet : 0.7320954907161804
Amyl : 0.8143236074270557
Benzos : 0.6923076923076923
Caff : 0.9734748010610079
Canabis : 0.7161803713527851
Choc : 0.9814323607427056
Coke : 0.6896551724137931
Crack : 0.896551724137931
Ecstasy : 0.7029177718832891
Heroin : 0.8938992042440318
Ketamin : 0.830238726790451
Legalh : 0.7427055702917772
LSD : 0.753315649867374
Meth : 0.8010610079575596
Mushrooms : 0.7241379310344828
Nicotine : 0.649867374005305
Semer : 0.9946949602122016
VSA : 0.8938992042440318
```



Random Forest

WHY : Random forests are robust and can handle multiple data types without prior normalization. They can capture complex interactions between features and are less sensitive to overfitting.

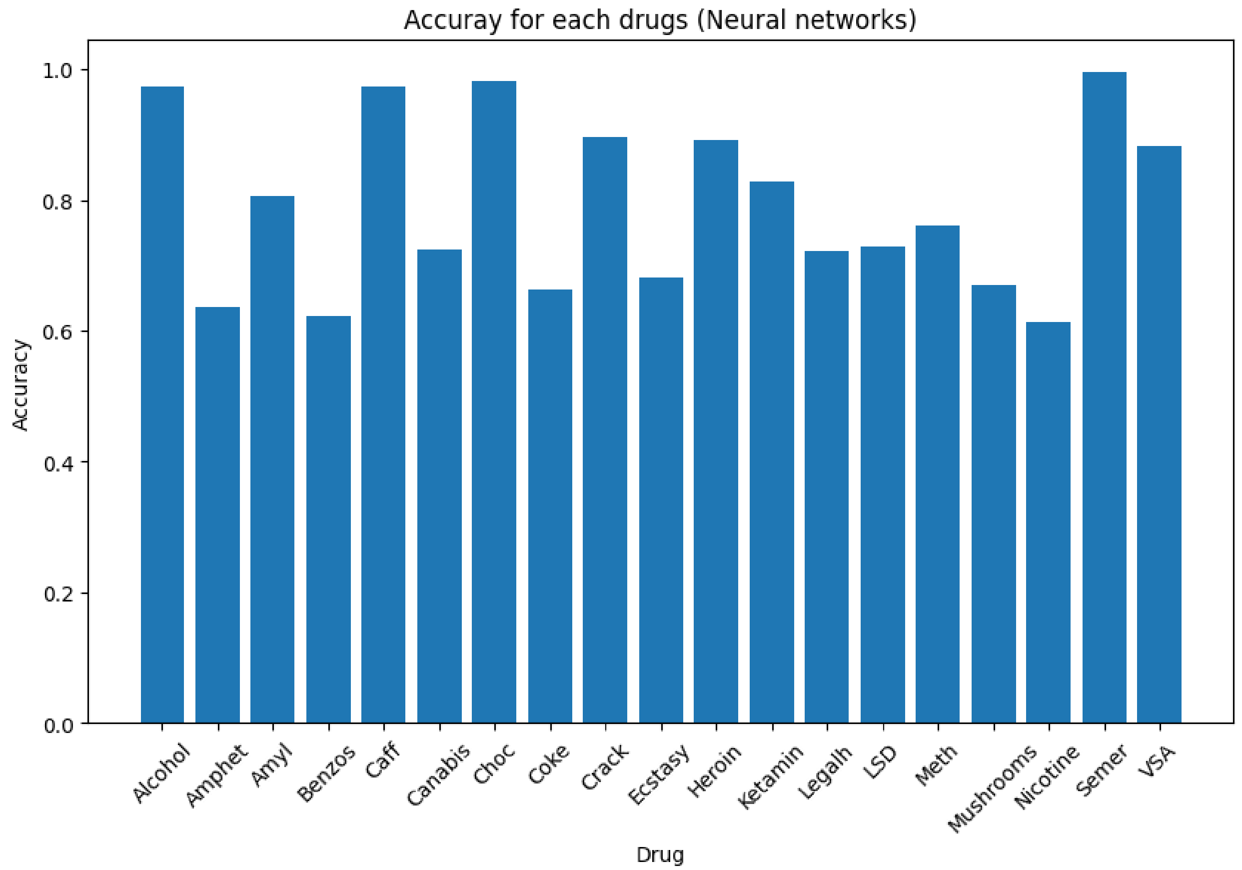
```
Alcohol : 0.9734748010610079
Amphet : 0.636604774535809
Amyl : 0.8063660477453581
Benzos : 0.623342175066313
Caff : 0.9734748010610079
Canabis : 0.7241379310344828
Choc : 0.9814323607427056
Coke : 0.6631299734748011
Crack : 0.896551724137931
Ecstasy : 0.6816976127320955
Heroin : 0.8912466843501327
Ketamin : 0.8275862068965517
Legalh : 0.7214854111405835
LSD : 0.7294429708222812
Meth : 0.7612732095490716
Mushrooms : 0.6710875331564987
Nicotine : 0.6127320954907162
Semer : 0.9946949602122016
VSA : 0.883289124668435
```



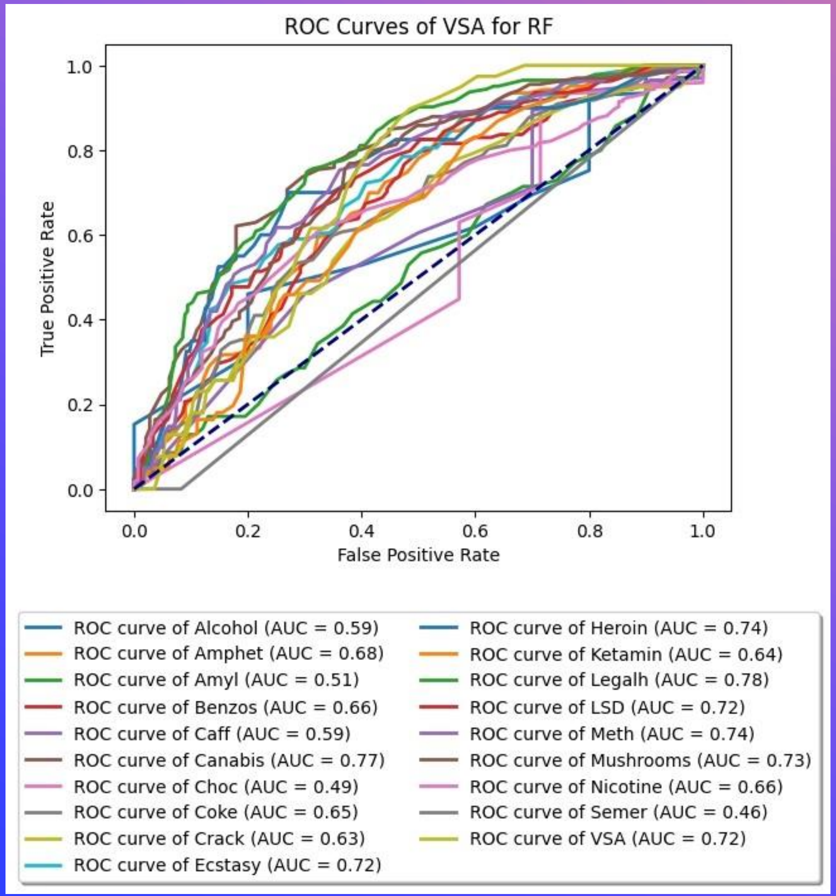
Neural Network

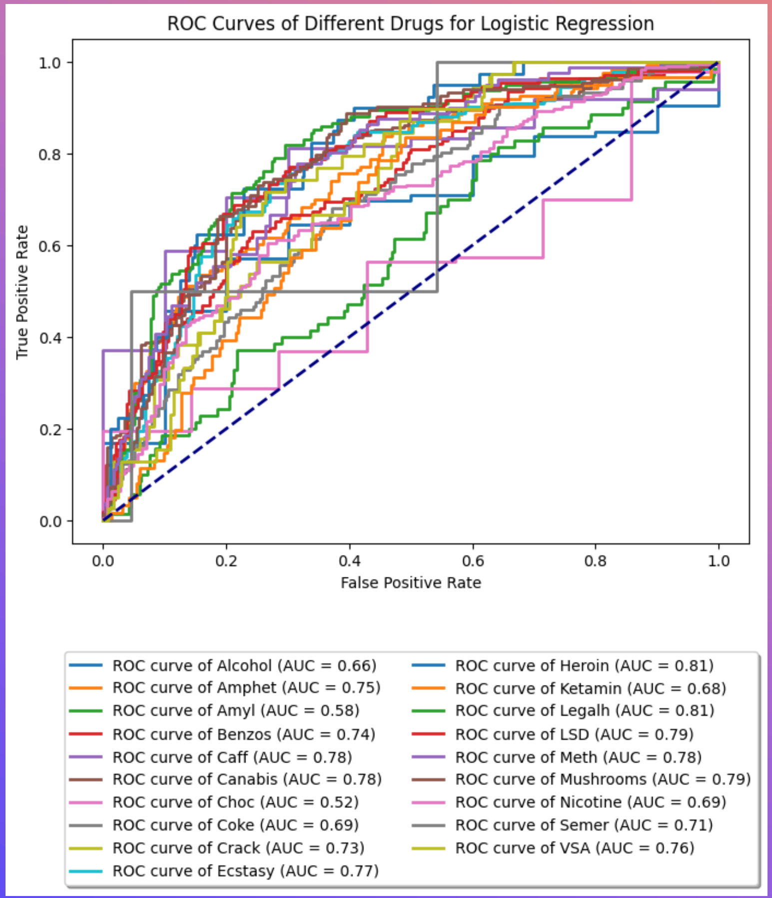
WHY : Neural networks can capture complex nonlinear patterns in data. They can be effective for learning subtle relationships between personality characteristics and drug use.

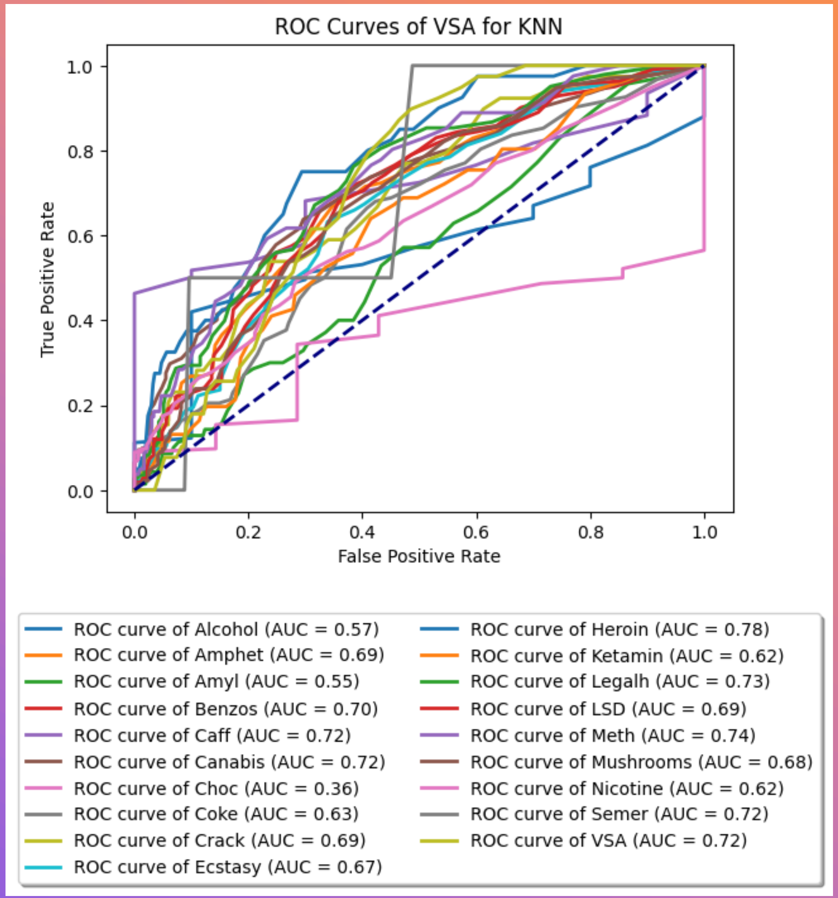
```
Alcohol : 0.9734748005867004
Amphet : 0.7161803841590881
Amyl : 0.8143236041069031
Benzos : 0.6870026588439941
Caff : 0.9734748005867004
Canabis : 0.73209547996521
Choc : 0.9814323782920837
Coke : 0.6896551847457886
Crack : 0.8965517282485962
Ecstasy : 0.6896551847457886
Heroin : 0.8938992023468018
Ketamin : 0.8381962776184082
Legalh : 0.7506631016731262
LSD : 0.7718833088874817
Meth : 0.7798408269882202
Mushrooms : 0.7427055835723877
Nicotine : 0.663129985332489
Semer : 0.9946949481964111
VSA : 0.8965517282485962
```

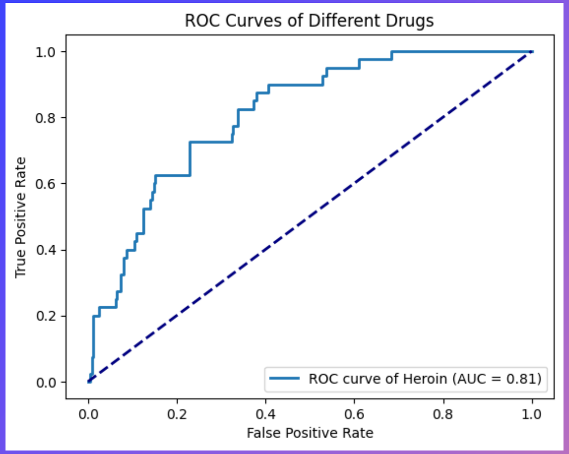


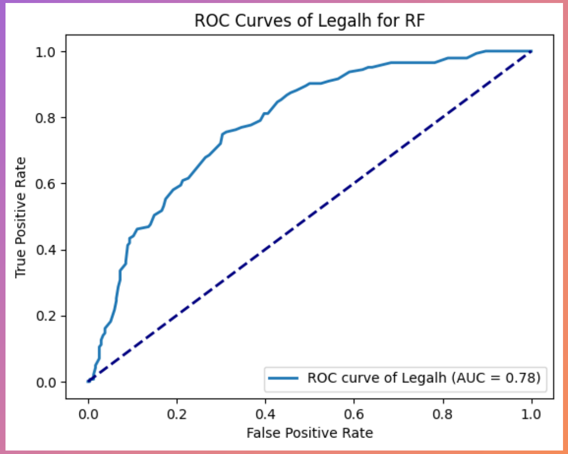
We traced the ROC curve which is the plot of false positives and true positives

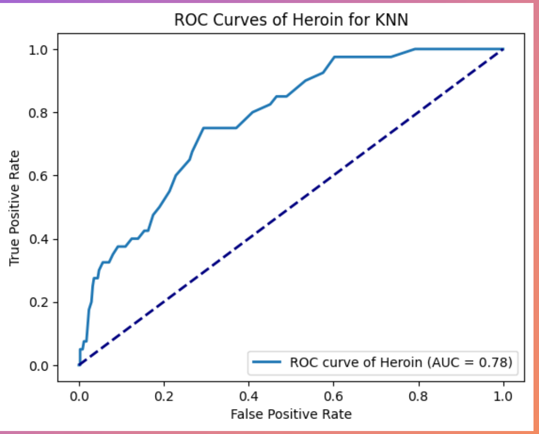




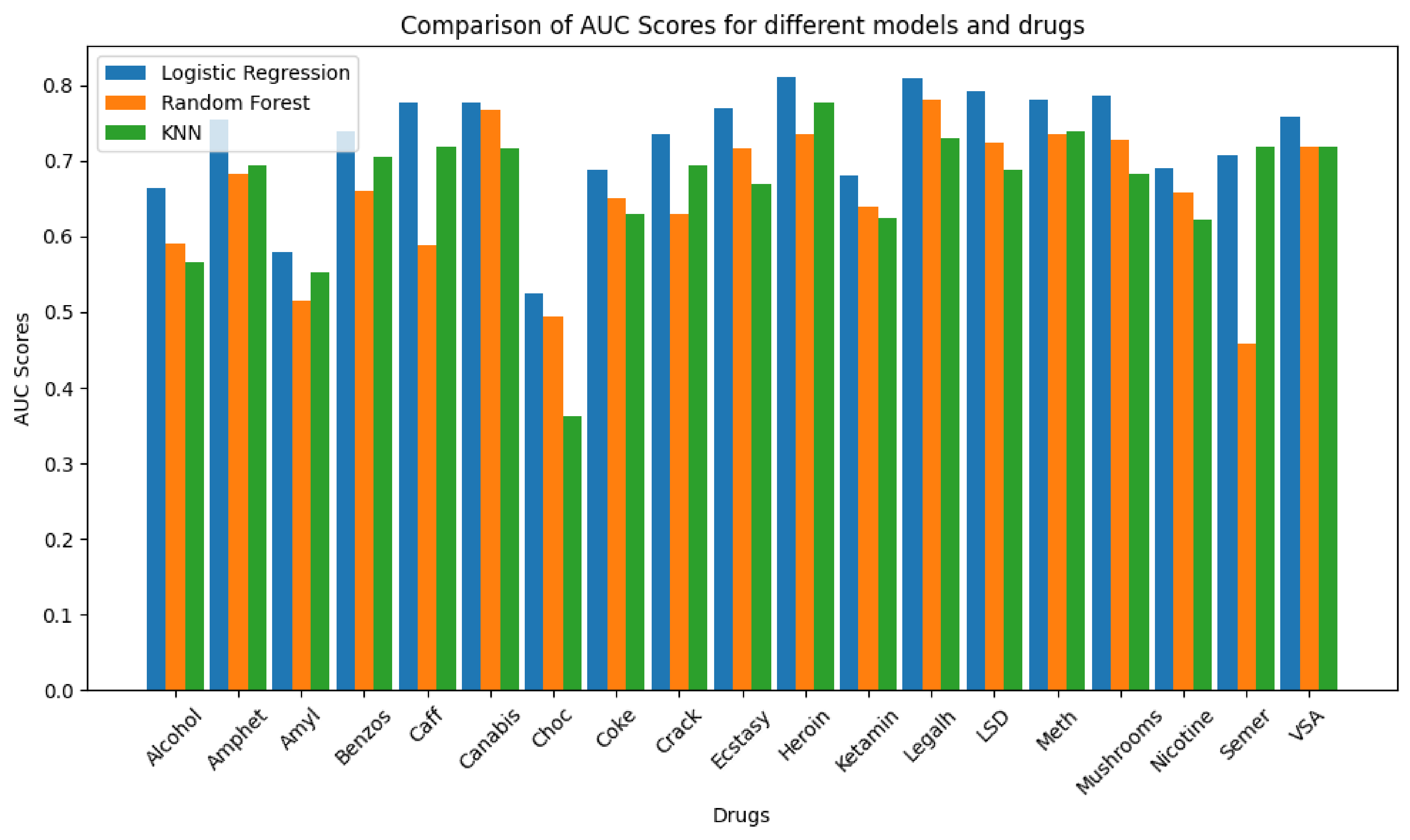








We then plotted them separately to make the graphs more readable.



To conclude on the models, we observe that Logistic regression is indeed a better model than Random Forest and KNN, because it has a higher AUC score, for our dataset. On the other hand, between Random Forest and KNN, we cannot choose the car that depends on drugs.

https://clipchamp.com/watch/UPsvSRpFsDd