



DRUG CONSUMPTION

# TABLE OF CONTENTS

- I. The dataset
  - I. General Presentation
  - II. Variables
  - III. Problems
- II. Classification
  - I. Type
  - II. The best One
- III. Flask

# I - THE DATASET : DRUG CONSUMPTION

## I. General Presentation:

### UCI Machine Learning Repository: Drug consumption (quantified) Data Set

- A Social topic dataset from : 2016-10-17
- This dataset contains : 1885 instances. Therefore, we don't have enough data in order to work with the dataset and to apply different classification algorithms to it but we are going with it.
- Credits:
  - Elaine Fehrman
  - Vincent Egan
  - Evgeny M. Mirkes



# I - THE DATASET : DRUG CONSUMPTION

## II. Variables

We have 32 features in total. Nevertheless, you are going to see that we will analyse important features and do some correlations in order to remove some features that don't have importance for the classification. All input attributes are originally categorical and are quantified.

The participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug they had 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.

Now let's look at the proportion of some features and analyse them.

# VARIABLES

## Feature : Country

In this MapPlot by plotly, we can see an unbalanced distribution of the data, we can assume that the poll was done in the UK. Nevertheless, if someone from another country does the drug prediction it could bias the result. So we may drop this feature during the classification

Nombre de consommateur

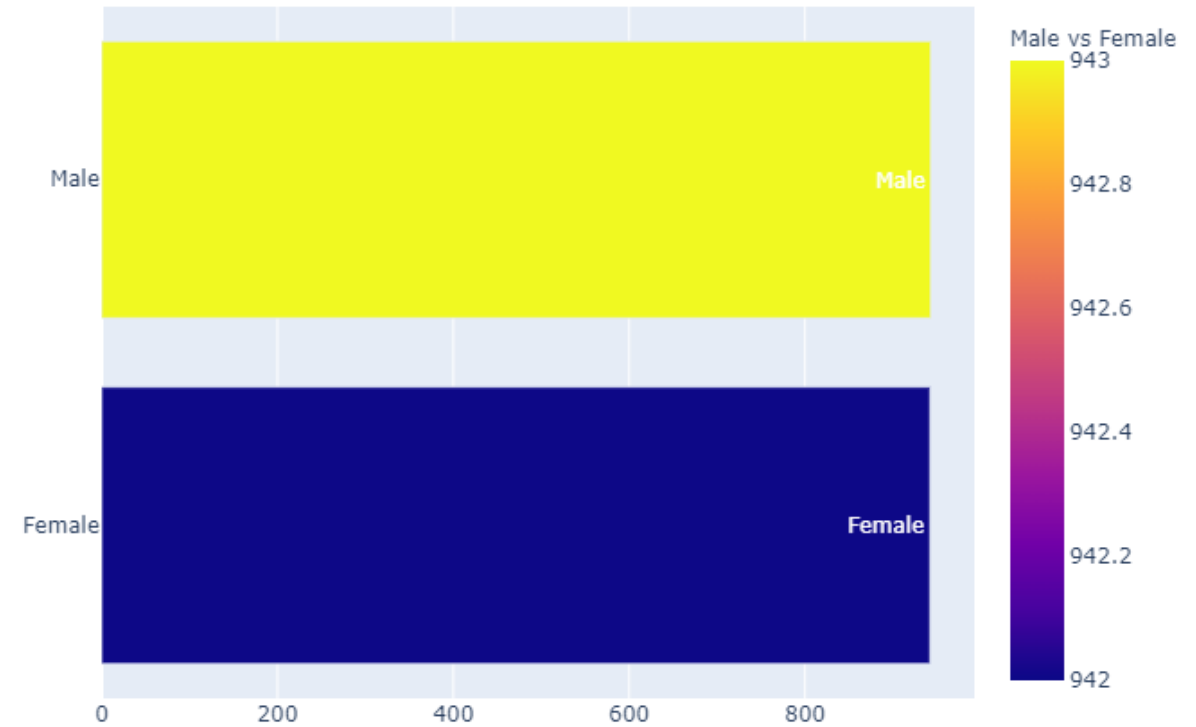


# VARIABLES

## Feature : Gender

There is an equal distribution of the gender.  
The gender will not be an issue during the classification.

Gender

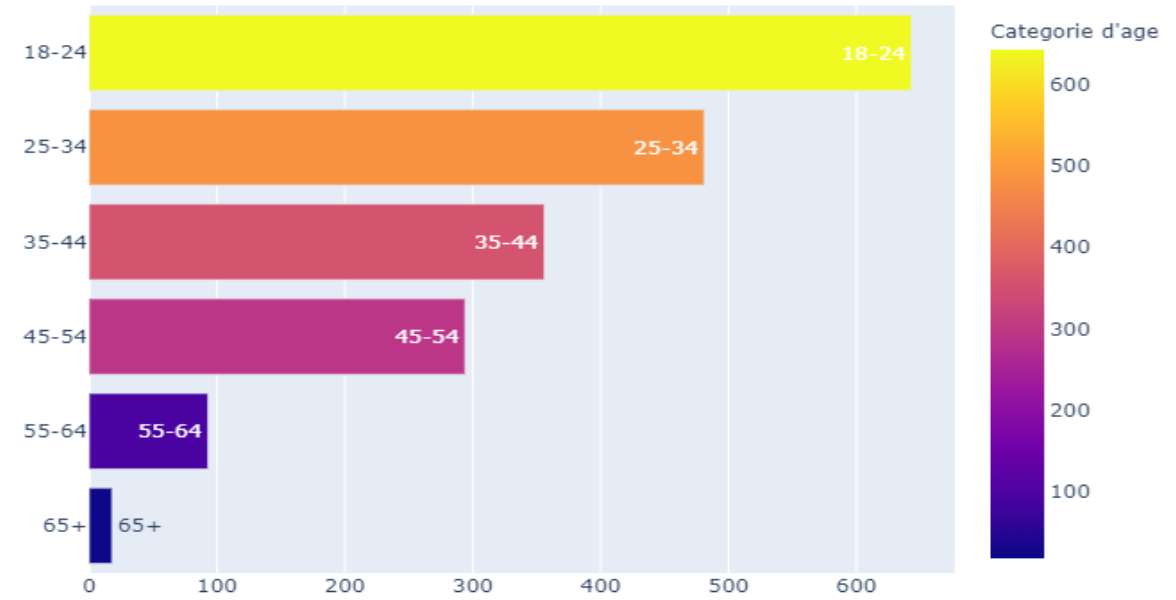


# VARIABLES

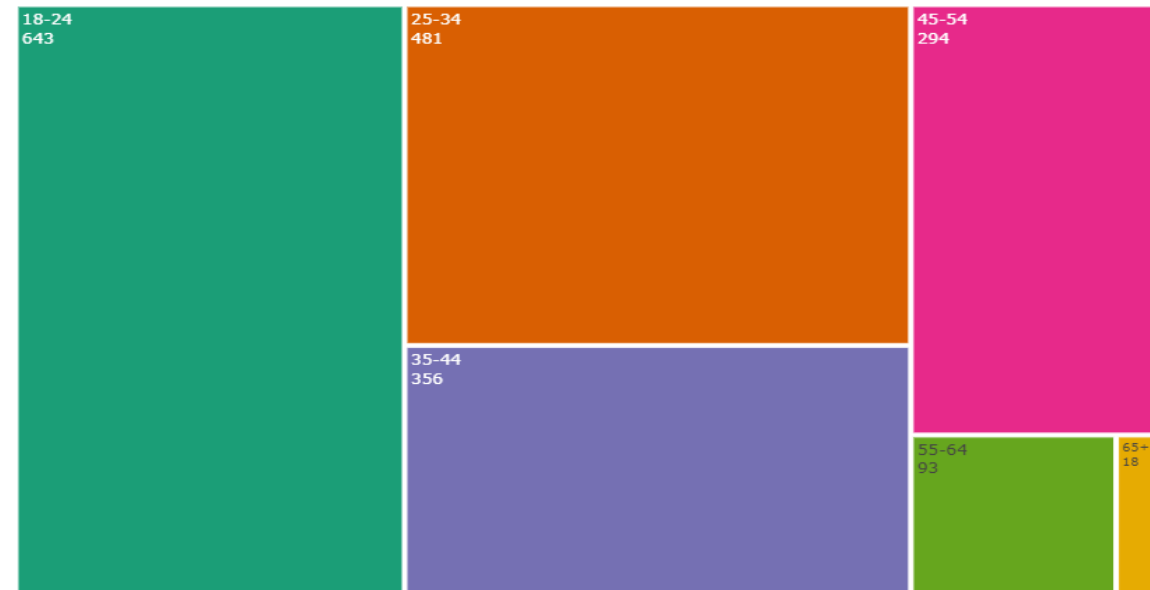
## Feature : Age

Here, we have a normal distribution. In fact, it's logical that there are more 18-24 drug users because we can speculate that they have a "you only live once" mindset that is pushing them to seek new sensations and try everything at least once.

Age



Categorie d'age

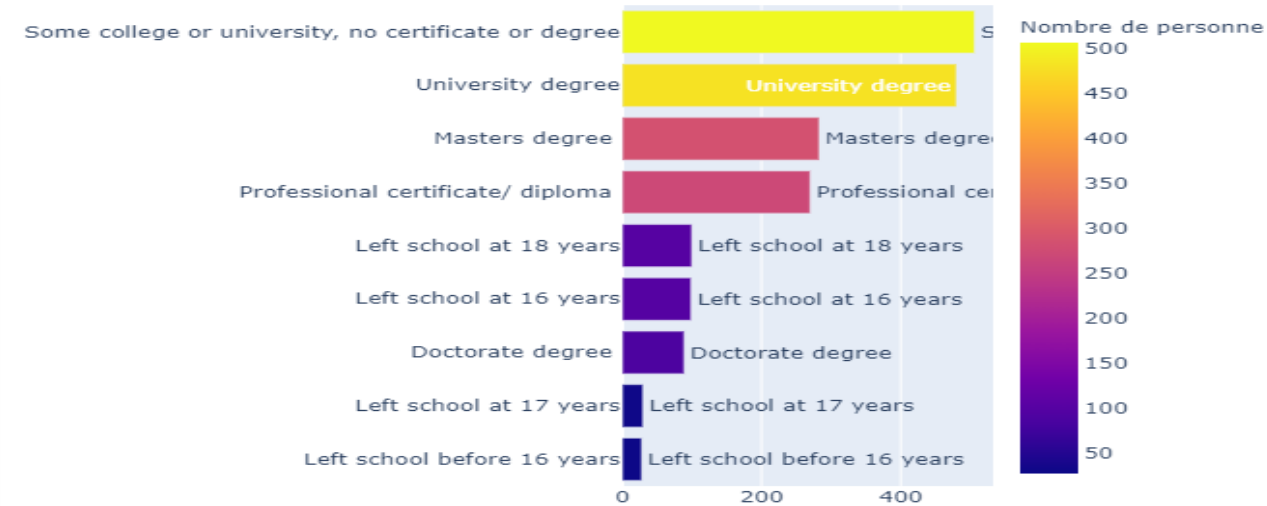


# VARIABLES

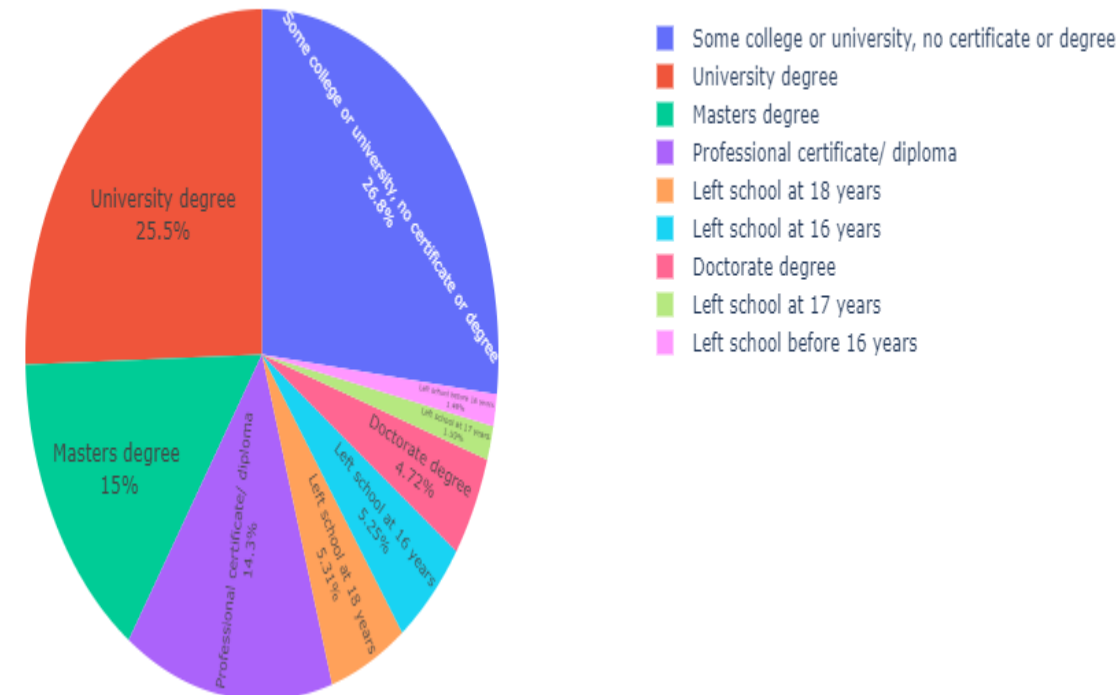
## Feature : Education

According to the feature Age, young people, so people in university or college, are more likely to use drugs because of the stress of school but also because some young people that don't have any degree or certificate suffered from stress thus use drugs to relax themselves.

### Education



Utilisateur de drogue selon leurs education



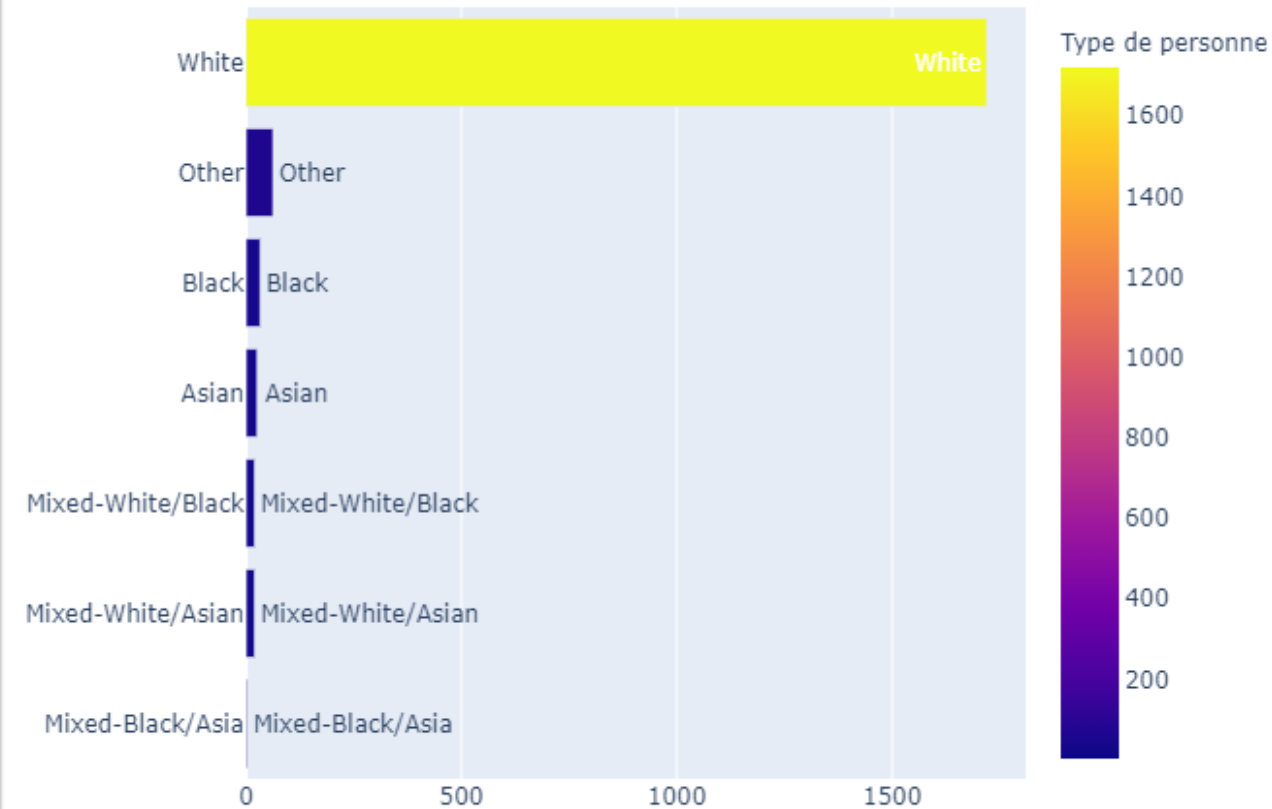


# VARIABLES

## Feature : Ethnicity

Here, we have a unbalanced feature. In fact, we can see that we have a large amount of white people in this dataset. Unfortunately, if we are from another ethnicity, the prediction will be most likely biased. We are going to drop this feature without any doubt and we are going to see it later thanks to a correlation graph.

Ethnicity

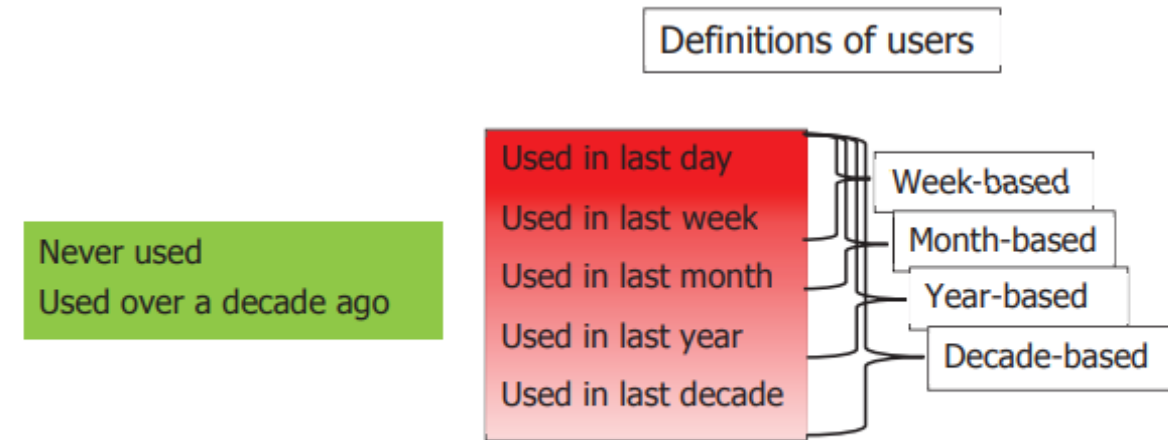


# VARIABLES

Feature : For each drug

➤ The categories ‘Used in last year’, ‘Used in last decade’, ‘Used over a decade ago’ and ‘Never used’ are combined to form a group of non-users and all three other categories are placed into the group of users. This classification problem is called ‘**month-based**’.

➤ We decided to use a “month-based” classification because for us the use of drug a decade ago or a year ago by someone might be solely for experimenting purposes therefore he might not be a real drug consumer. If we kept them as users, it might compromise our classification.



**Fig 1. Categories of drug users.** Categories with green background always correspond to drug non-users. Four different definitions of drug users are presented.

Source of the article [1]: <https://arxiv.org/pdf/1506.06297.pdf>

# I - THE DATASET : DRUG CONSUMPTION

## II. Variables

### Drug Groups

Given that we don't have enough data for each drug, we are going to do three big groups of drug according to : <https://arxiv.org/pdf/1506.06297.pdf>

- The Heroin pleiad (heroinPl) includes crack, cocaine, methadone, and heroin.
- The Ecstasy pleiad (ecstasyPl) includes amphetamines, cannabis, cocaine, ketamine, LSD, magic mushrooms, legal highs, and ecstasy.
- The Benzodiazepines pleiad (benzoPl) includes contains methadone, amphetamines, and cocaine.

# I - THE DATASET : DRUG CONSUMPTION

## II. Variables

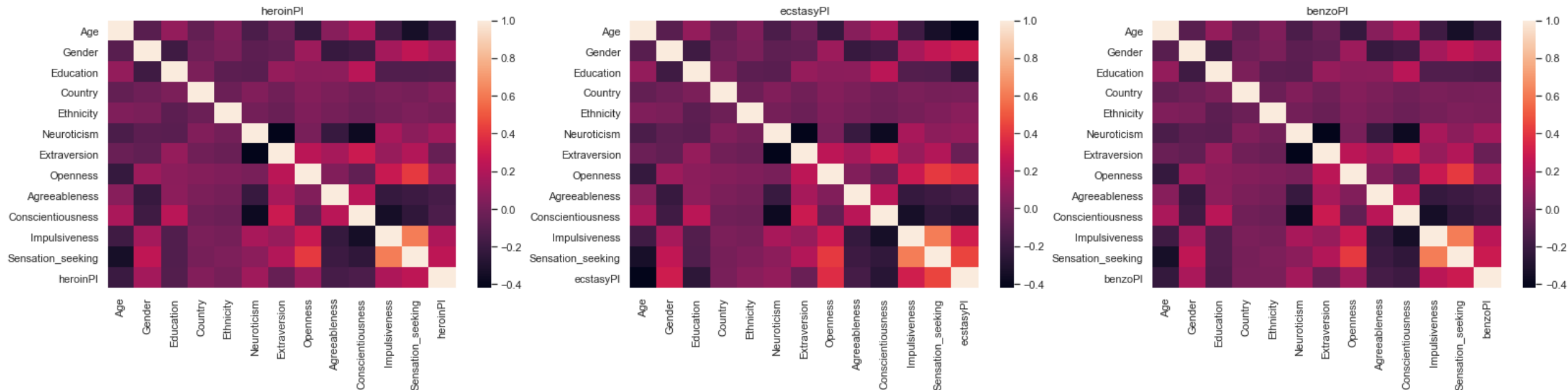
After analysing some features, we decided to create correlation for the three drugs. In order to do that, we have to decode some features such as Age, Gender, Education, Country, Ethnicity.

Age	Gender	Education	Country	Ethnicity
'18-24' age -> 0	Female -> 0	Left school before 16 years -> 0	Australia -> 0	Asian -> 0
'25-34' age -> 1	Male -> 1	Left school at 16 years -> 1	Canada -> 1	Black -> 1
'35-44' age -> 2		Left school at 17 years-> 2	New Zealand->2	Mixed-Black/Asian -> 2
'45-54' age -> 3		Left school at 18 years-> 3	Other -> 3	Mixed-White/Asian -> 3
'55-64' age -> 4		Some college or university, no	Republic of Ireland ->4	Mixed-White/Black -> 4
'65+' age -> 5		certificate or degree -> 4	UK ->5	Other -> 5
		Professional certificate/ diploma -> 5	USA ->6	White -> 6
		University degree -> 6		
		Masters degree -> 7		
		Doctorate degree -> 8		

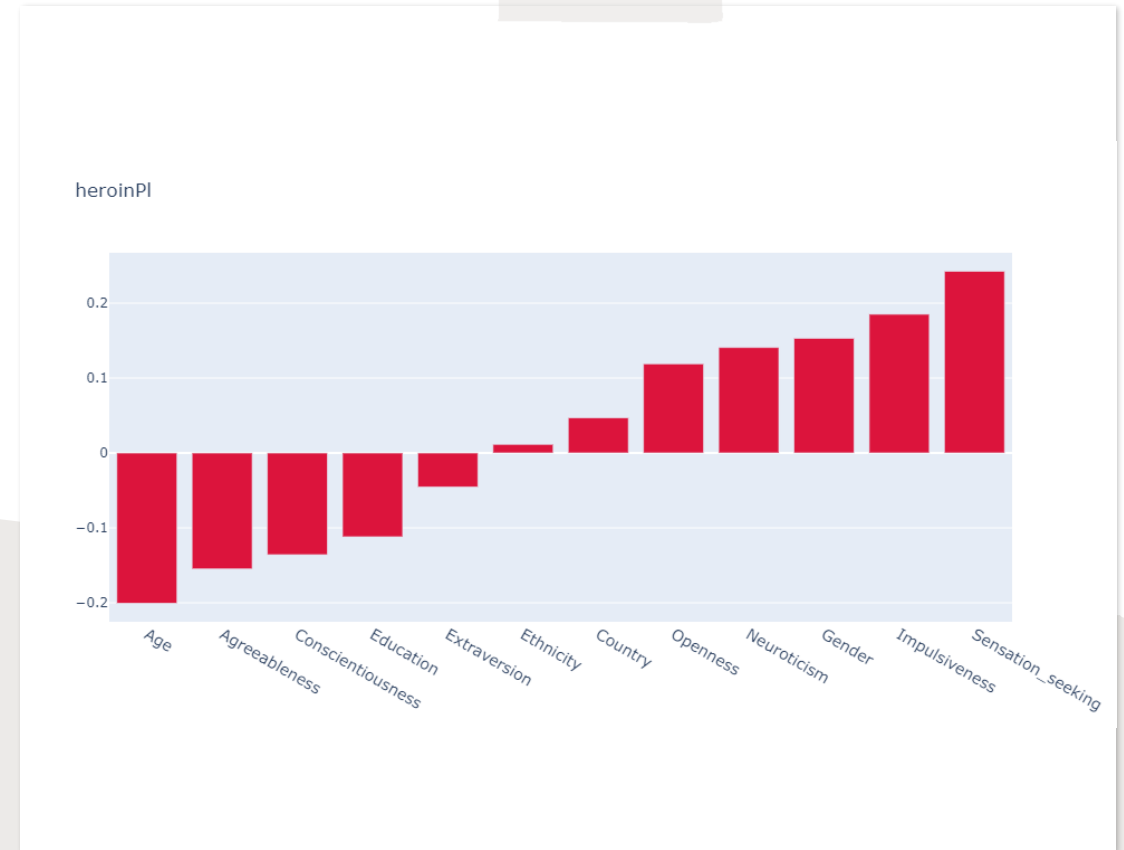
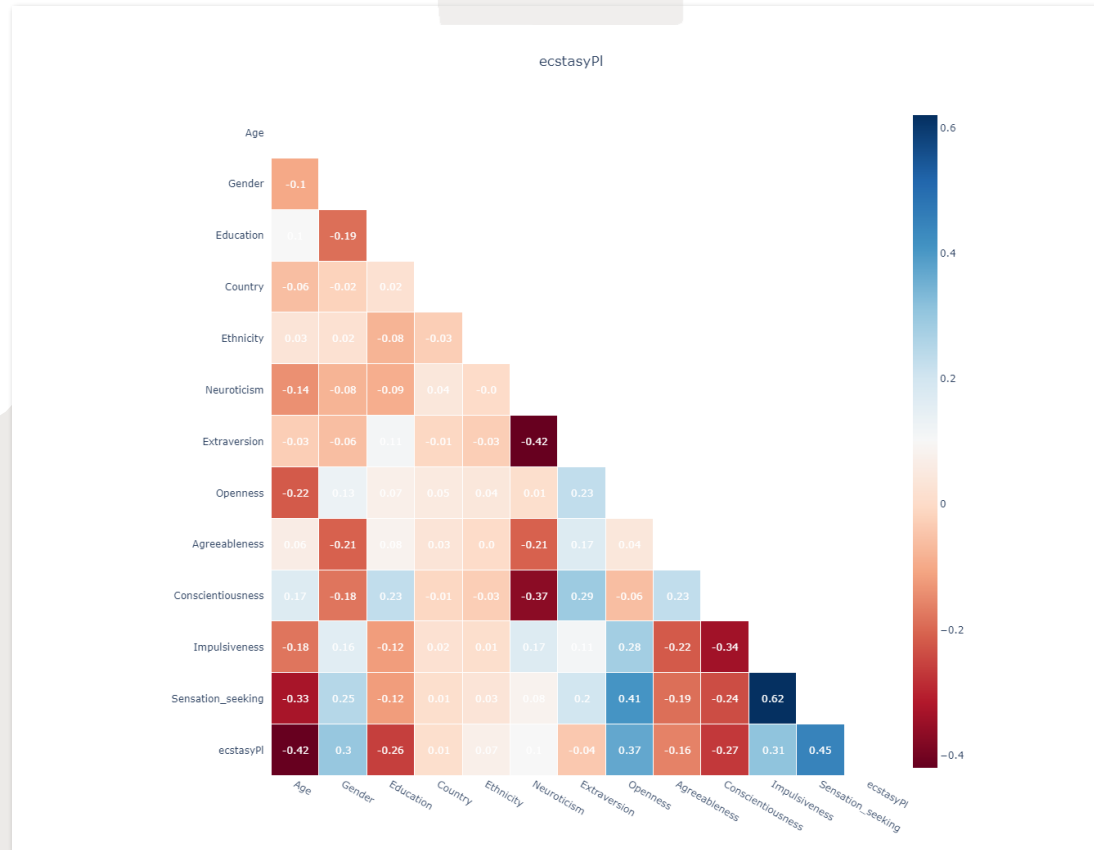
# I - THE DATASET : DRUG CONSUMPTION

## II. Variables

Here is a global vision of the correlation of the features with each drugs. We are going to explore this more precisely for each drug groups.





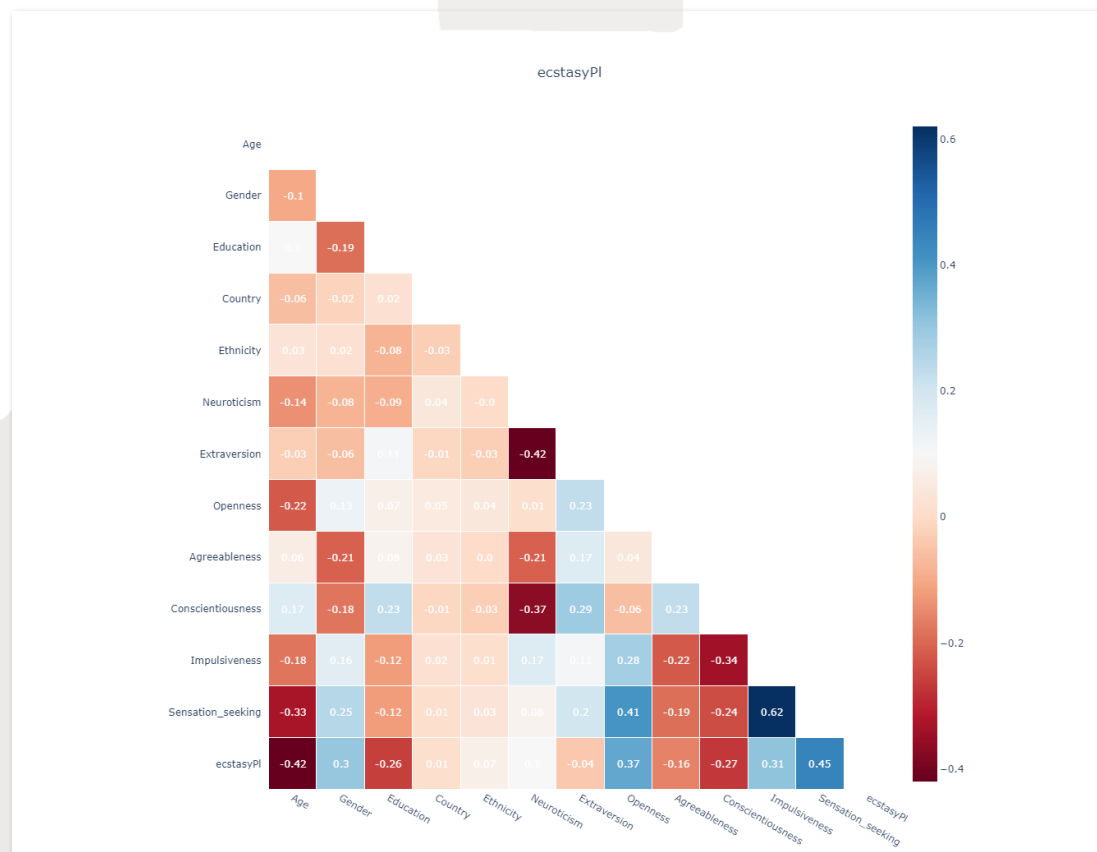


# CORRELATION

## HeroinPL

The correlation of each feature is under 0.5 because there are not enough data and each features are badly distributed.

As we saw during the analyse of each features, the features Country and Ethnicity had no importance in the drugs feature, in fact their correlation is near 0. It means that they don't play a role in the attribution of drug user or not user. Furthermore, we can also drop the extraversion feature that has a small negative correlation which is not enough to influence the prediction.

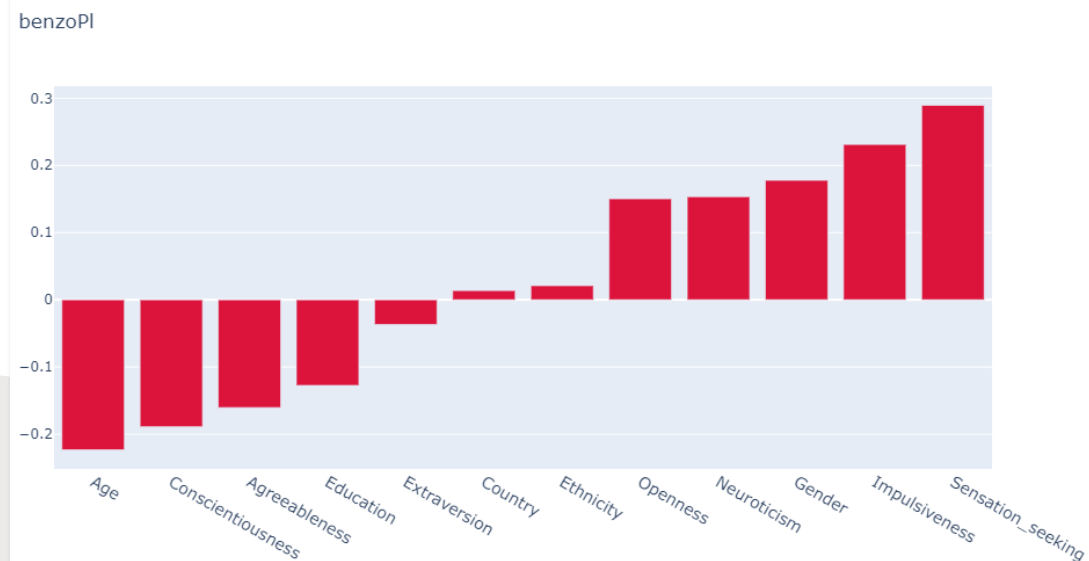
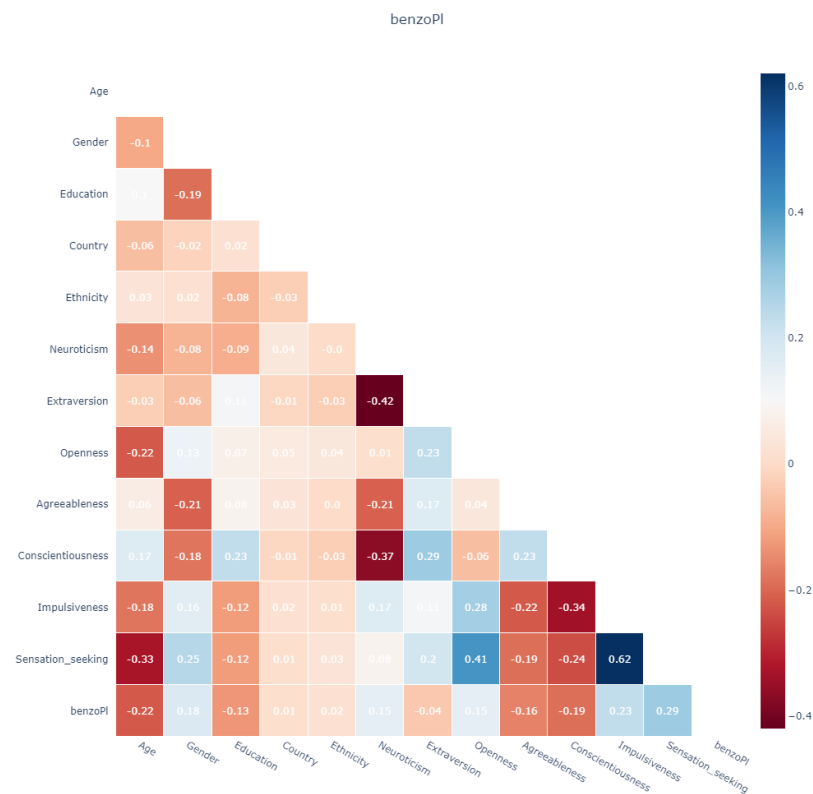


# CORRELATION

## EcstasyPL

The correlation of each feature is under 0.5 because there are not enough data and each features are badly distributed.

Here we have the same case as the HeroinPI drugs.



# CORRELATION

## BenzoPL

The correlation of each feature is under 0.5 because there are not enough data and each features are badly distributed.

Here we have the same case as the HeroinPl and EcstasyPl drugs.

# I - THE DATASET : DRUG CONSUMPTION

## III. Problems

Through the first part, we resolved and explained to you some problems such as :

- We simplified the classifications by regrouping the drugs in three groups that we explained in the first part.
- We transformed to binary classification by union of part of classes into one new class for monthly user. For example, "Never Used", "Used over a Decade Ago", "Used over a Year Ago" form class "Non-user" and all other classes form class "User". The best binarization of classes for each attribute.

1 - if a person used a drug in month, week or day, then he did consume a drug.

0 - other categories are placed into the group that he did not consume a drug.

- We tried to find the best binarization of classes for each attribute.

Now, we are going to solve another problem, even if we don't have enough data, we are going to evaluate the risk to be a drug consumer for each drug. This will be our modelling part.

## II-CLASSIFICATION: PREDICTION

We are going to predict if someone is a user or not of one of the three drugs: HeroinPl, EcstasyPl, BenzoPl. Thus we are going to use three classification Machine Learning Algorithms. For this, we used scikit-learn library as we did in personal work during the class.

### HOW DO WE PROCEED ?

First of all, we create three dataframes which correspond to the three drugs then we divide it in training set and test\_set. Then, we scaled the train and test set. We create an algorithm inspired by the PW5 which tests all the classification algorithms from sklearn library with their default parameters and we keep the top 15 which have the best average. Then, we test the first three or the three that we know the theoretical.

However, at the end of the project, when we predict with using the best algorithm with the best average with feature that we know it's a drug user we don't have the expected result. After some research on internet, we noticed that we didn't take into account the good metric which is F1-score. So we redo some algorithm.



# II-CLASSIFICATION: PREDICTION

## HOW DO WE PROCEED ?

In fact, for unbalanced class we have to use the F1-score. In our case, HeroinPl and BenzoPl are unbalanced classes.

```
data_heroin['heroinPl'].value_counts()
```

```
0    1576  
1     309  
Name: heroinPl, dtype: int64
```

```
data_ecstasy['ecstasyPl'].value_counts()
```

```
0     951  
1     934  
Name: ecstasyPl, dtype: int64
```

```
data_benzo['benzoPl'].value_counts()
```

```
0    1463  
1     422  
Name: benzoPl, dtype: int64
```



Unbalanced

## II-CLASSIFICATION: PREDICTION

### HOW WE PROCEED ?

#### Confusion Matrix:

		Predicted Class	
		0	1
True Class	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

#### F1-Score:

The f1-score is a subtle mix between the average and the sensibility. It's more interesting than average because the true negative is not taken into account. That is very useful in unbalanced situation. The closer it is to 1 the better is the prediction.

$$\text{F1-Score} = 2 \frac{(\text{tp}/(\text{tp}+\text{fp})) * (\text{tp}/(\text{tp}+\text{fn}))}{\text{tp}/(\text{tp}+\text{fp}) + \text{tp}/(\text{tp}+\text{fn})}$$

*[ML : Précision, F1-Score, Courbe ROC, que choisir ? / by Beranger Natanelic / Medium](#)*

## II-CLASSIFICATION: PREDICTION

### HOW DO WE PROCEED ?

#### *Fine-Tuning:*

We are now entering into the fine tuning part. What we mean by fine-tuning is rearranging the algorithm in order for it to fit our dataset better than before. Unfortunately, for the Ecstasy dataset and Benzopl dataset , we fine-tuned first with the best average algorithm which are Logistic Regression , Random Forest Classifier and K-nn and then for the F1-score.

HeroinPl		EcstasyPl	BenzoPl	
Average	F1	Logistic Regression() RandomForestClassifier() KNeighborsClassifier()	Average	F1
Logistic Regression() RandomForestClassifier() KNeighborsClassifier()	NearestCentroid() Perceptron() BernouilliNB()		Logistic Regression() RandomForestClassifier() KNeighborsClassifier()	NearestCentroid() ExtraTreeClassifier() GaussianNB()

## II-CLASSIFICATION: PREDICTION

### HOW DO WE PROCEED ?

#### *Fine-Tuning:*

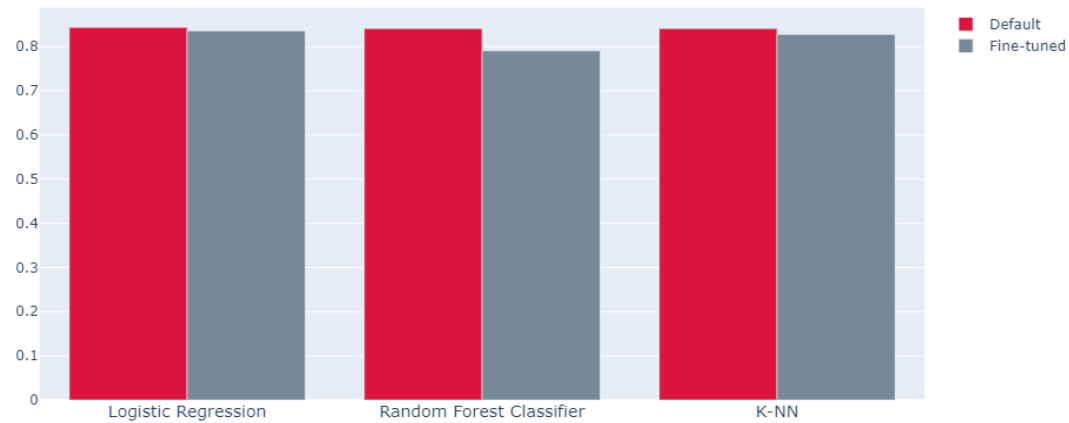
In order to find the best hyperparameter and the best combination, we create an algorithm which permits to do a gridsearch such as we saw in PW. However, when we fine-tuned each algorithm, either we had the same average(or f1-score) or a lower score which surprised us a little.

```
def test_hyperparametres(algo, hyperparametres,a, b):  
    grid      = GridSearchCV(algo, hyperparametres, n_jobs=-1)  
    grid.fit(a, b)  
    print (grid.best_score_, grid.best_estimator_)  
    return grid.best_score_, grid.best_estimator_
```

In order to find the best between the fine tuned one and the default one, we do a bar chart for each drug classes.

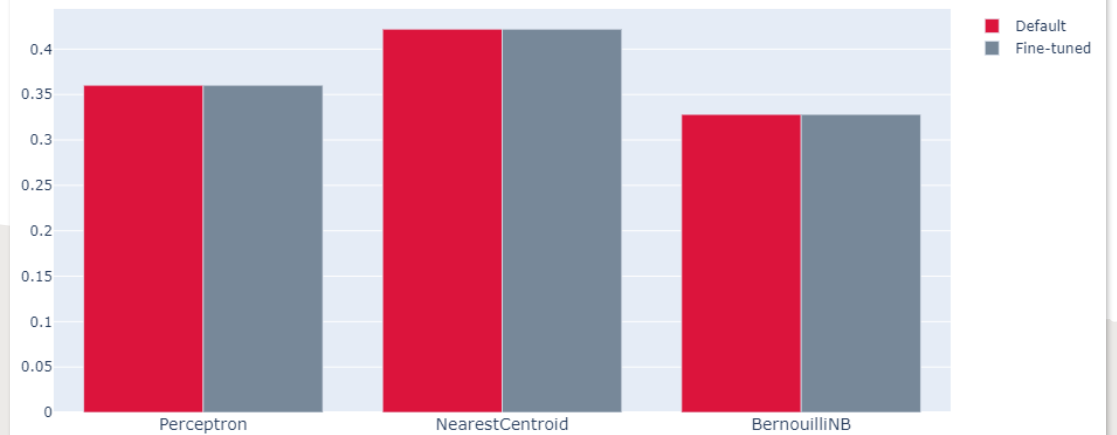
## AVERAGE

HeroinPI



## F1 - SCORE

HeroinPI



## FINAL RESULT:

HeroinPL

The F1-score is under 0.5 because first we don't have enough data and the correlation of the feature is too low as we saw in the first part.

We choose the NearestCentroid Algorithm to predict this class of drug.

```
NearestCentroid().get_params()  
{'metric': 'euclidean', 'shrink_threshold': None}
```



# FINAL RESULT:

## EcstasyPL

The F1-score is under 0.5 because first we don't have enough data and the correlation of the feature is too low as we saw in the first part.

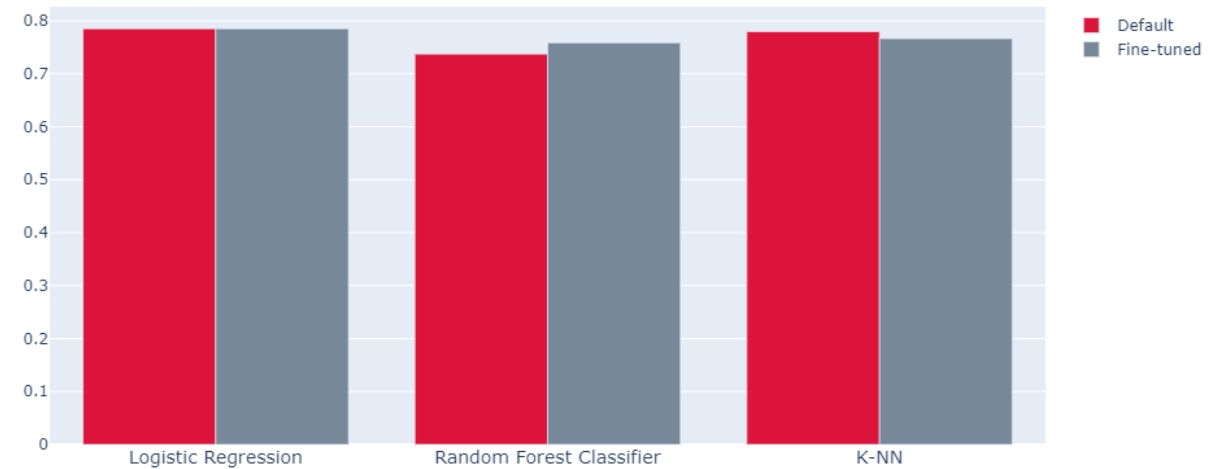
We choose the `LogisticRegression()` algorithm to predict this class of drug

```
a=test_hyperparametres(lr, param_grid,X2_train, y2_train)
```

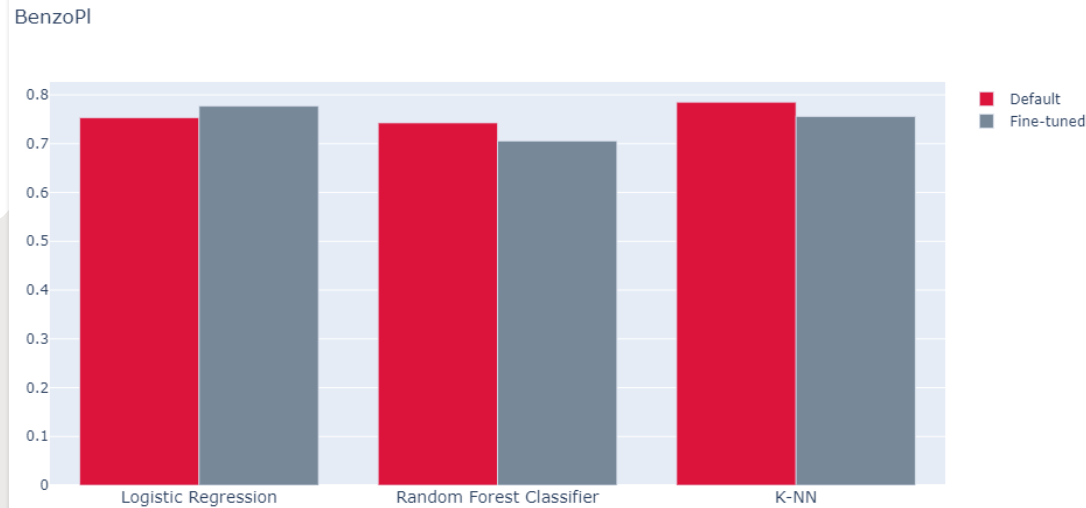
```
0.7877868473740952 LogisticRegression(C=0.1, class_weight='balanced', solver='liblinear')
```

## AVERAGE

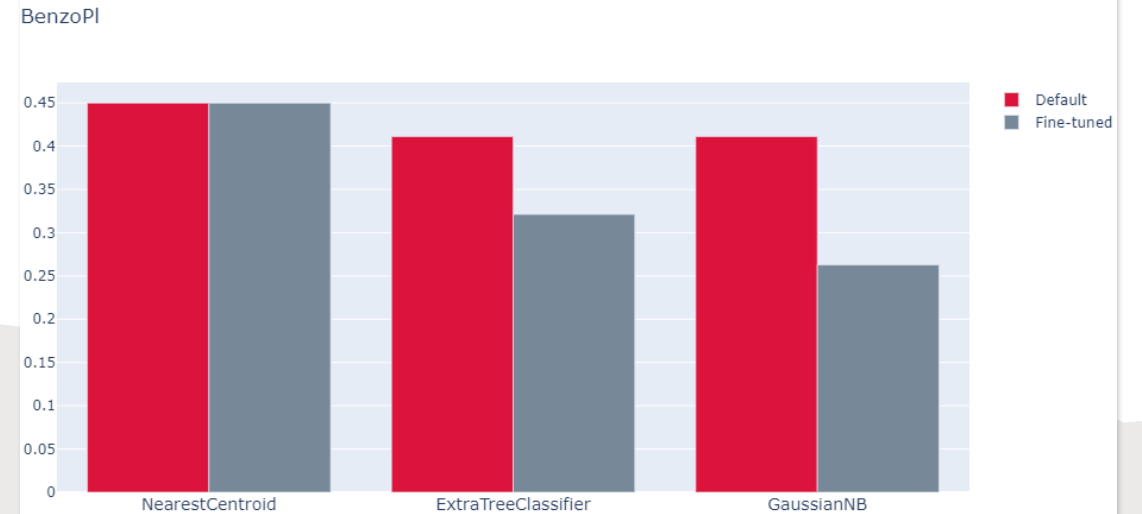
EcstasyPL



## AVERAGE



## F1 - SCORE



## FINAL RESULT:

BenzoPL

The F1-score is under 0.5 because first we don't have enough data and the correlation of the feature is too low as we saw in the first part.

We choose the NearestCentroid Algorithm to predict this class of drug.

```
test_hyperparametersf1(nc, param_grid,X3_train, y3_train)  
NearestCentroid(shrink_threshold=0.0)
```

### III - FLASK

Concerning the API, we found that it could be interesting for the visitor to be able to enter his features, and that our model predicts if this user has a drug profile or not.

For this, we used the flask module.

# API FOLDER ARCHITECTURE

Nom	Modifié le	Type	Taille
static	05/01/2022 18:45	Dossier de fichiers	
templates	05/01/2022 16:01	Dossier de fichiers	
drug_consumption.data	06/12/2021 22:17	Fichier DATA	339 Ko
app.py	05/01/2022 18:37	Fichier PY	5 Ko
model.py	03/01/2022 22:43	Fichier PY	7 Ko

Static : contains 2 script .js | 9 images | 3 fonts

Templates : contains 3 html files

Drug\_consumption.data : dataset

App.py : script that launches flask

Model.py : script that train the dataset and can take features in parameters and return the probability of a profile user to be a regular drug user or not

# A P I

With the help of a raspberry pi, we host the website and you can access it wherever and whenever you want via this link :

[www.trademanager.fr:5000](http://www.trademanager.fr:5000)



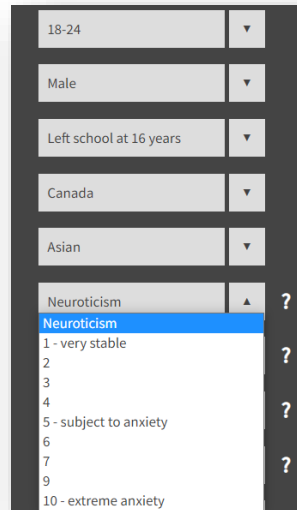
WELCOME TO THE DRUG USER DETECTOR

Age	▼	
Gender	▼	
Education	▼	
Country	▼	
Ethnicity	▼	
Neuroticism	▼	?
Extraversion	▼	?
Openness	▼	?
Agreeableness	▼	?
Conscientiousness	▼	?
Impulsiveness	▼	?
Sensation_seeking	▼	?
<b>SUBMIT</b>		



# A P I

You can complete your features :

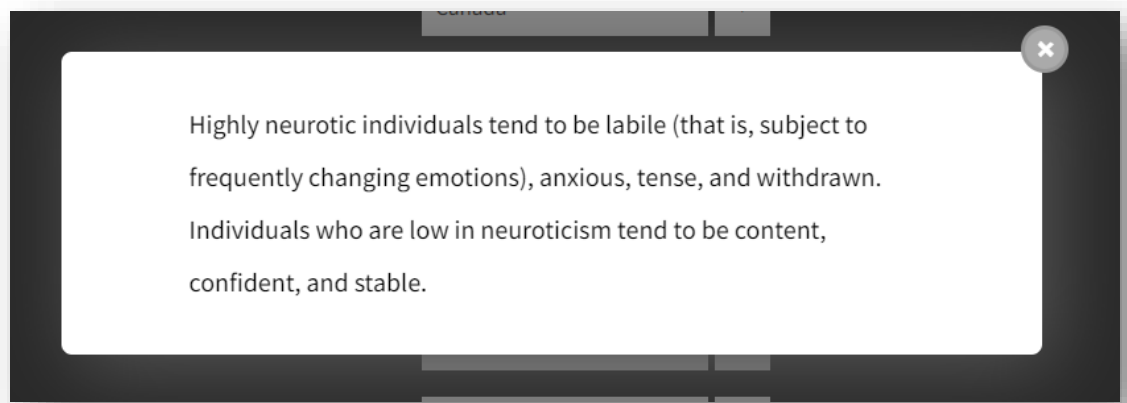


18-24	▼	
Male	▼	
Left school at 16 years	▼	
Canada	▼	
Asian	▼	
Neuroticism	▲	?
Neuroticism		
1 - very stable		?
2		
3		
4		
5 - subject to anxiety		?
6		
7		
9		
10 - extreme anxiety		?

If you need more information about a feature, you can click the tool-tip button :



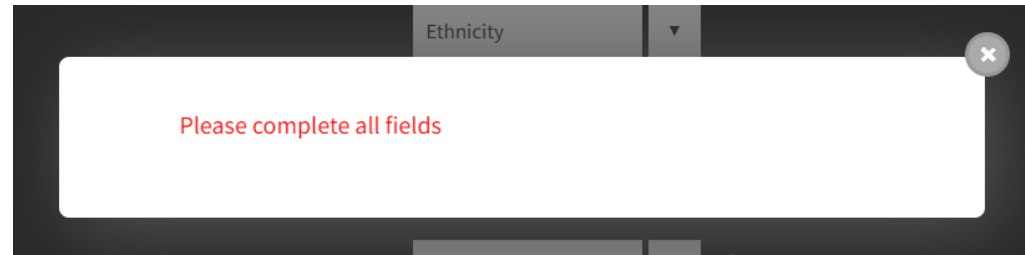
Neuroticism	▼	?
-------------	---	---



Highly neurotic individuals tend to be labile (that is, subject to frequently changing emotions), anxious, tense, and withdrawn. Individuals who are low in neuroticism tend to be content, confident, and stable.

# A P I

Make sure to complete all features, if you don't, you will not be able to access the results

A screenshot of a web form interface. At the top, there is a dark grey header bar with the word "Ethnicity" in white text and a small downward arrow icon. Below the header is a large white rectangular input area. Inside this area, the text "Please complete all fields" is displayed in red. In the top right corner of the white area, there is a small grey circular button with a white "x" icon, likely for closing a modal or notification.

When it's done, click on the submit button :



# RESULTS

## RESULTS

You are likely to be a Heroin user at 55 % ?

You are likely to be an Ecstasy user at 18 % ?

You are likely to be a Benzo user at 56 % ?

[More informations about the dataset](#)

[More informations about the model](#)

## INTERPRETING THE RESULTS

You have three categories of drugs with a probability to be a regular drug user for each one.

When the probability is higher than 55 %, the color is in red and it seems like you have a profile of regular user for this drug.

We must consider that the F1-score is not high thus we don't need to have a blind trust according to the benzoPl and heroinPl predictions because of the lack of data. That's why we take 55% and not 50% for HeroinPl and EcstasyPl.

You can click on the tool-tip button for each one to know the details of a category.

## MORE INFORMATIONS

If you want to, you can access some images describing the dataset and the models by clicking on these buttons :

