Drug Consumption

Python for data analysis
Arthur VINOT and Alexis SOK

The data

- IN : The dataset is composed with different informations of people such as : ID, Age Gender, Education, Country, Ethnicity and personality.
- OUT : Moreover we have information about their drugs consumption such as alcohol, amphet, amyl, benzos. These features have values between 0 to 6 to evaluate their habits (never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day).

The problems

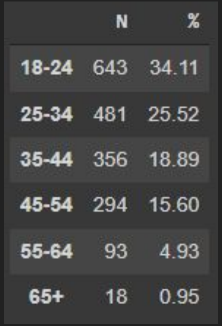
This dataset is really rich and a lot of problem could be resolved, for instance :

- Predict if someone will have an addiction to drugs
- Predict which drugs they will be consum
- Predict for each drug, if someone will try it or be a drug addict

Data Cleaning

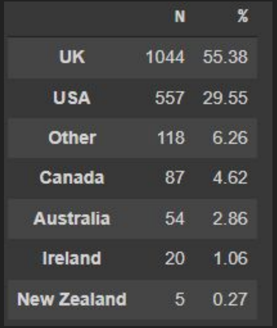
- Our data has no missing values
- For the drugs, their values are strings so we convert them into numerical values
- Moreover, we had to map values of age, gender, ethnicities, countries, educations because their were numerical values that have no signification

Data exploration and analysis - people analysis



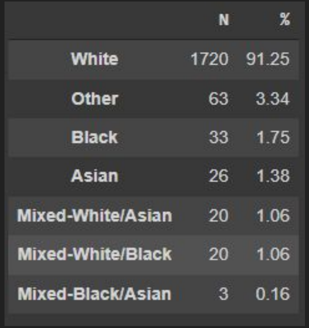
	N	%
18-24	643	34.11
25-34	481	25.52
35-44	356	18.89
45-54	294	15.60
55-64	93	4.93
65+	18	0.95

Age range



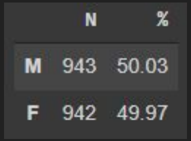
	N	%
UK	1044	55.38
USA	557	29.55
Other	118	6.26
Canada	87	4.62
Australia	54	2.86
Ireland	20	1.06
New Zealand	5	0.27

Countries



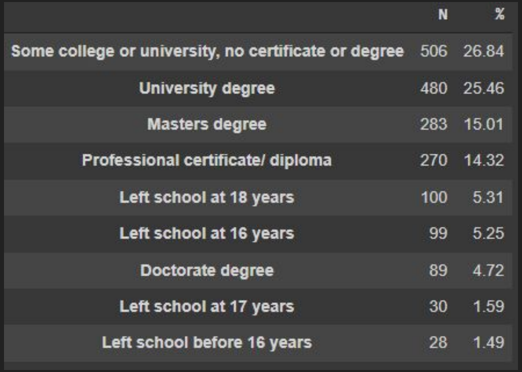
	N	%
White	1720	91.25
Other	63	3.34
Black	33	1.75
Asian	26	1.38
Mixed-White/Asian	20	1.06
Mixed-White/Black	20	1.06
Mixed-Black/Asian	3	0.16

Ethnicities



	N	%
M	943	50.03
F	942	49.97

Gender



	N	%
Some college or university, no certificate or degree	506	26.84
University degree	480	25.46
Masters degree	283	15.01
Professional certificate/ diploma	270	14.32
Left school at 18 years	100	5.31
Left school at 16 years	99	5.25
Doctorate degree	89	4.72
Left school at 17 years	30	1.59
Left school before 16 years	28	1.49

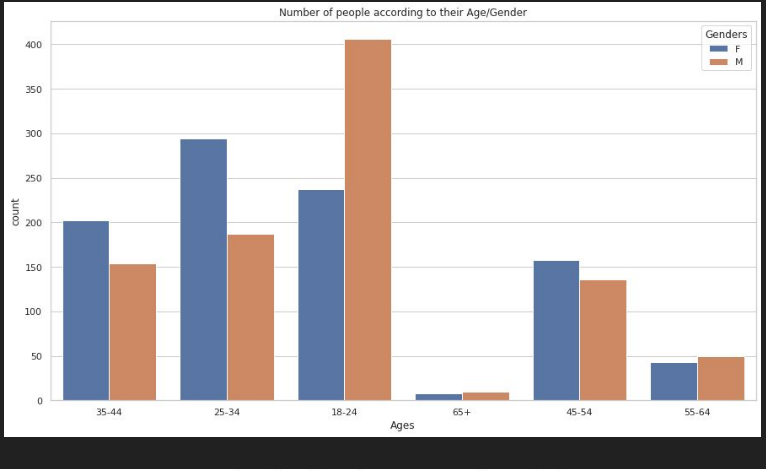
Educations

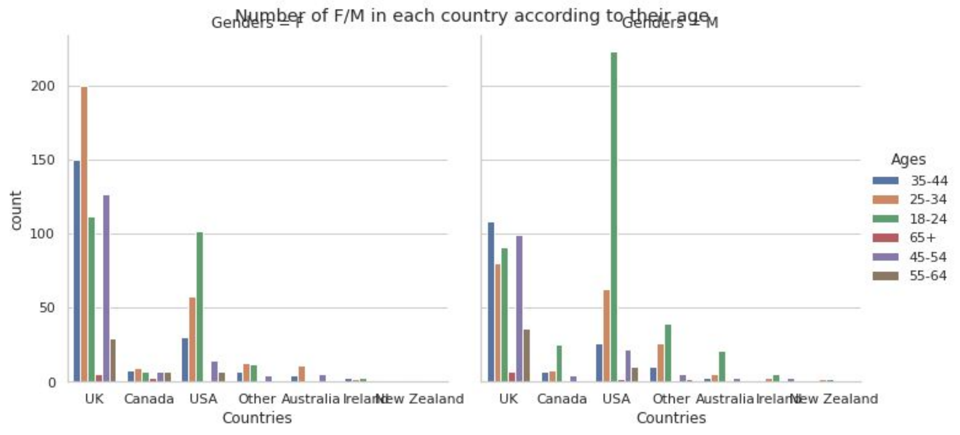
Our dataset is composed by 50% male and 50% female.

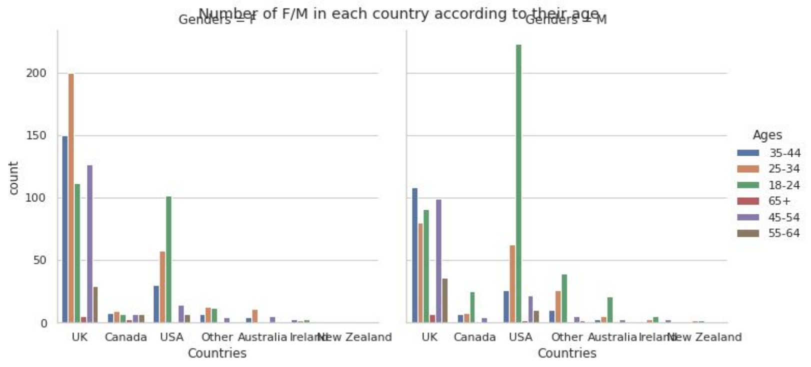
Most of people are white and they came from english speaking countries.

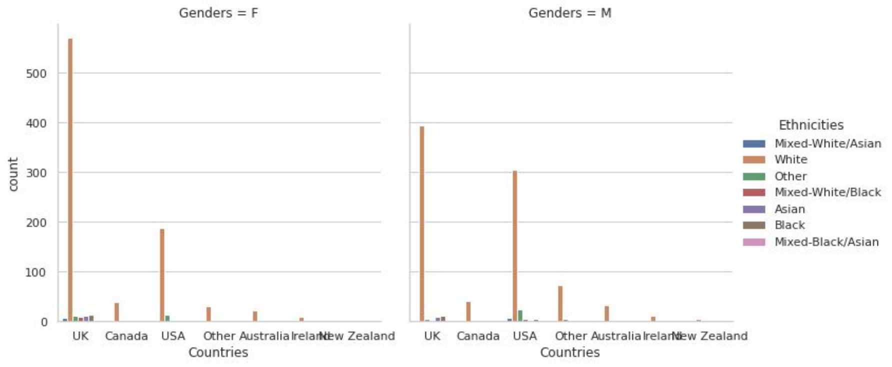
Remarks : Beside gender, we think that the rests of the features are unbalanced and not representative.

Data exploration and analysis - people analysis

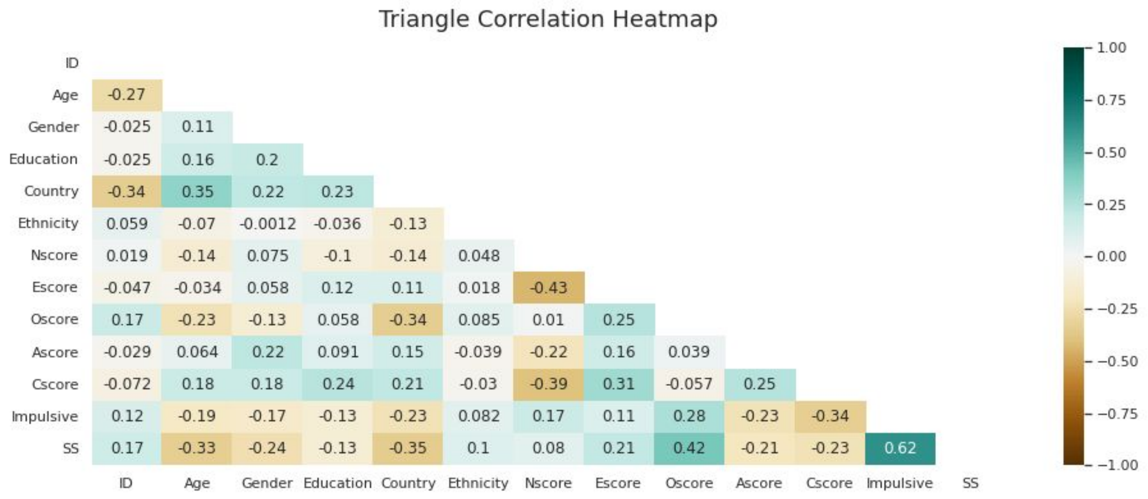








Data analysis - Correlation



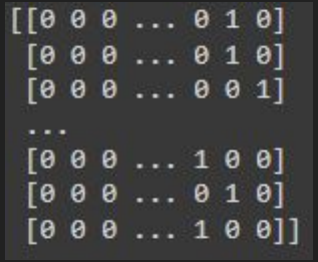
Triangle Correlation Heatmap

	ID	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS
ID													
Age	-0.27												
Gender	-0.025	0.11											
Education	-0.025	0.16	0.2										
Country	-0.34	0.35	0.22	0.23									
Ethnicity	0.059	-0.07	-0.0012	-0.036	-0.13								
Nscore	0.019	-0.14	0.075	-0.1	-0.14	0.048							
Escore	-0.047	-0.034	0.058	0.12	0.11	0.018	-0.43						
Oscore	0.17	-0.23	-0.13	0.058	-0.34	0.085	0.01	0.25					
Ascore	-0.029	0.064	0.22	0.091	0.15	-0.039	-0.22	0.16	0.039				
Cscore	-0.072	0.18	0.18	0.24	0.21	-0.03	-0.39	0.31	-0.057	0.25			
Impulsive	0.12	-0.19	-0.17	-0.13	-0.23	0.082	0.17	0.11	0.28	-0.23	-0.34		
SS	0.17	-0.33	-0.24	-0.13	-0.35	0.1	0.08	0.21	0.42	-0.21	-0.23	0.62	

As we can see the Sensation and the impulsiveness are correlated

Machine learning - data spliting

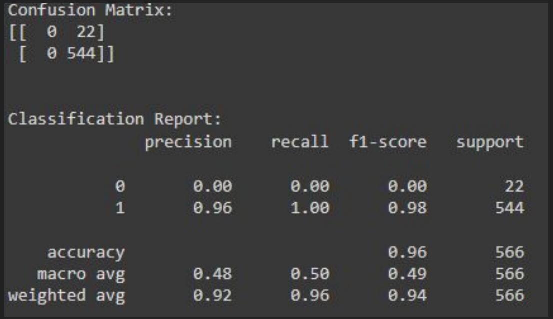
- For the data prediction we use the alcohol consumption : so we create a matrix 7x7 that evaluate with a 1 their alcohol consumption from the first index to the 7th index e.g. :
  These are going to be our actual values



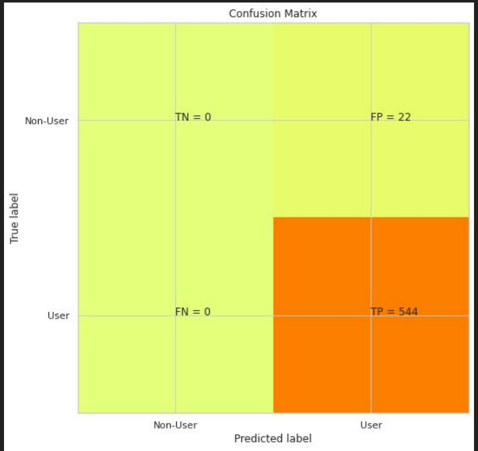
```
[[0 0 0 ... 0 1 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 1 0 0]]
```

- We scale our data and split them into 70% for the training set and 30% for the test set
- For each model we created function that evaluate our model : confusion matrix, precision, recall, score …

Machine learning - SVM



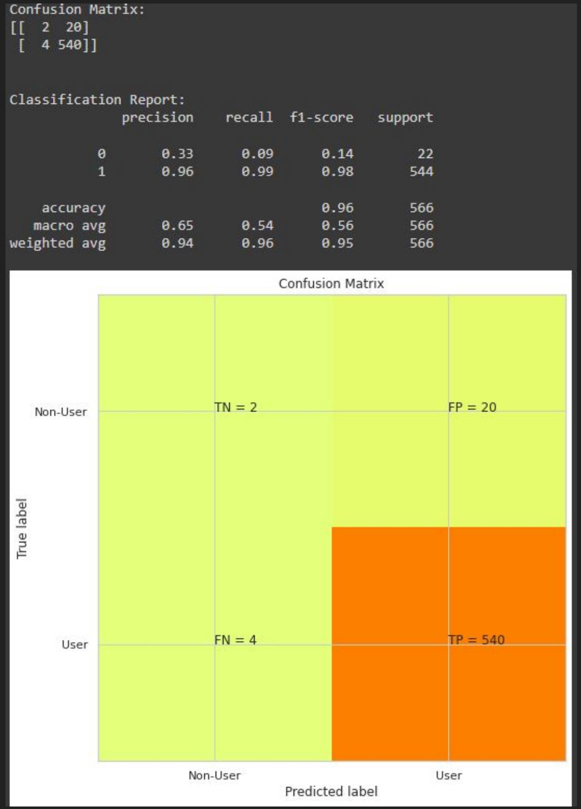
```
Confusion Matrix:
[[  0  22]
 [  0 544]]


Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        22
           1       0.96      1.00      0.98       544

    accuracy                           0.96       566
   macro avg       0.48      0.50      0.49       566
weighted avg       0.92      0.96      0.94       566
```



Observations :

- This model has a good score and precision = 0.98
- Among 566 observation, 544 were True

Machine learning - Decision Tree

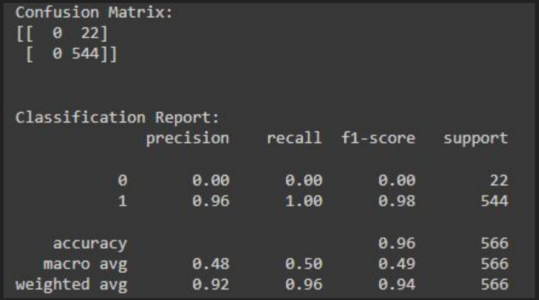


```
Confusion Matrix:
[[  2  20]
 [  4 540]]


Classification Report:
              precision    recall  f1-score   support

           0       0.33      0.09      0.14        22
           1       0.96      0.99      0.98       544

    accuracy                           0.96       566
   macro avg       0.65      0.54      0.56       566
weighted avg       0.94      0.96      0.95       566
```

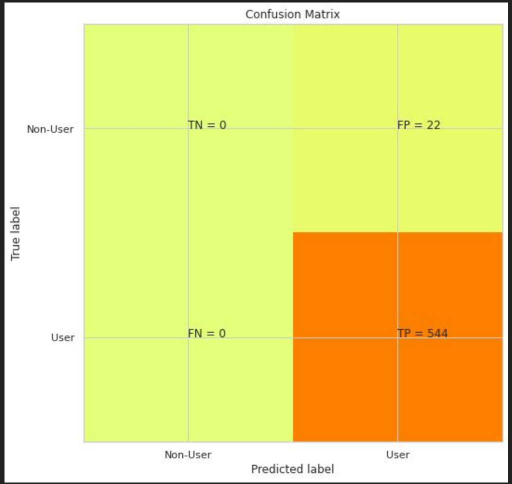
Observations :

- In comparison to the SVM, this one is less precise and accurate.

Machine learning - KNN



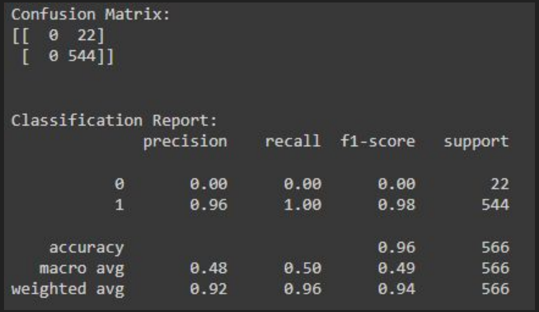
```
Confusion Matrix:
[[  0  22]
 [  0 544]]


Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        22
           1       0.96      1.00      0.98       544

    accuracy                           0.96       566
   macro avg       0.48      0.50      0.49       566
weighted avg       0.92      0.96      0.94       566
```



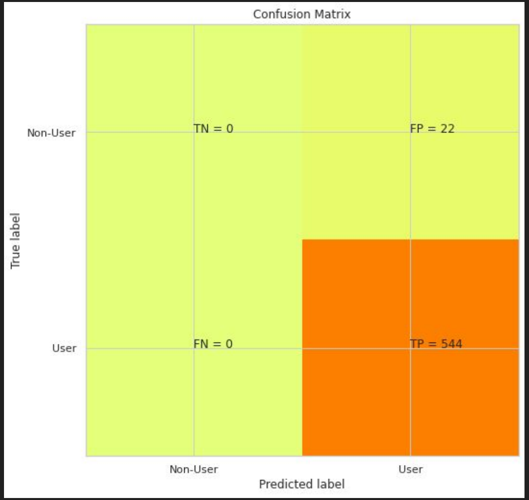
Observations :

- We tried the different values for the neighbors (2,5,6,7).
- Our best value were 5 because over this value the result were the same.
- This model is the same to SVM

Machine learning - RandomForest



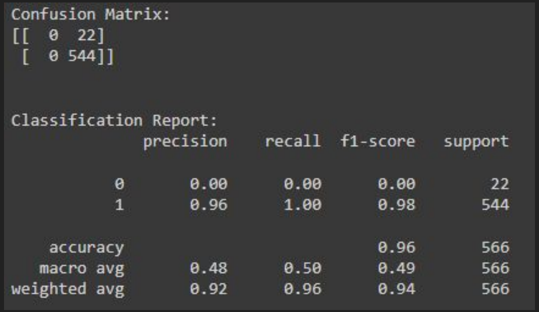
```
Confusion Matrix:
[[  0  22]
 [  0 544]]


Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        22
           1       0.96      1.00      0.98       544

    accuracy                           0.96       566
   macro avg       0.48      0.50      0.49       566
weighted avg       0.92      0.96      0.94       566
```



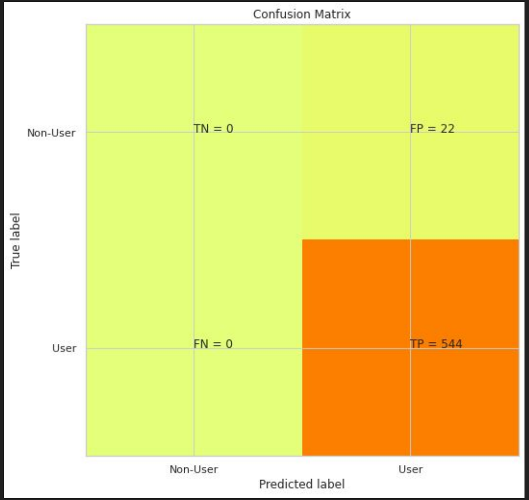
Observations :

- This model is the same to SVM

Machine learning - Logistic Regression



```
Confusion Matrix:
[[  0  22]
 [  0 544]]


Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        22
           1       0.96      1.00      0.98       544

    accuracy                           0.96       566
   macro avg       0.48      0.50      0.49       566
weighted avg       0.92      0.96      0.94       566
```



Observations :

- This model is the same to previous

Machine learning - models

- Every model were good and gave the same result except for decision tree.
- All of them were quiet quick less than 2 seconds

Conclusion of the project

This project was really interesting because we could practice everything we learnt in your courses and also in other courses such as Machine Learning.

This kind of project is the proof that data is a real treasure and we can make relevant prediction that could surprise us.