Projet Python for Data Analysis : Drug Consumption Analysis

Par Amato Jean-Emmanuel et Boika Emmanuel

Plan

- Problématique
- Approche de problème
- Solution envisagée
- Test
- Solution retenue
- Résultat

Problématique

I. Détails du dataset

Nous avons des données de 1885 répondants. 12 attributs sont connus : 7 attributs concernant les mesures de la personnalité : neuroticisme, extraversion, ouverture à l'expérience, agréabilité & conscience, impulsivité et recherche de sensations. Et 5 attributs concernant les données démographiques : le niveau d'éducation, l'âge, le sexe, le pays de résidence et l'ethnicité. Tous les attributs d'entrée sont initialement catégoriques et sont quantifiés. Après quantification, les valeurs de toutes les caractéristiques d'entrée peuvent être considérées comme des valeurs réelles. De plus, les participants ont été interrogés sur leur consommation de 18 drogues légales et illégales (alcool, amphétamines, nitrite d'amyle, benzodiazépine, cannabis, chocolat, cocaïne, caféine, crack, ecstasy, héroïne, kétamine, legal highs, LSD, méthadone, champignons, nicotine et substances volatiles) et d'une drogue fictive (Semeron) qui a été introduite pour identifier les surdéclarants. Pour chaque drogue, les participants doivent choisir l'une des réponses suivantes : ils n'ont jamais consommé la drogue, ils l'ont consommée il y a plus de dix ans ou au cours de la dernière décennie, de l'année, du mois, de la semaine ou du jour.

II. Informations sur les features

- ID est le numéro de l'enregistrement dans la base de données originale.
- Age est l'âge du participant et a une des valeurs suivantes :

Value	Meaning
-0.95197	18-24
-0.07854	25-34
0.49788	35-44
1.09449	45-54
1.82213	55-64
2.59171	65+

- Gender est le sexe du participant :

Value	Meaning
0.48246	Féminin
-0.48246	Masculin

- Education est le niveau d'éducation du participant et a une des valeurs :

Value	Meaning
-2.43591	A quitté l'école avant 16 ans
-1.73790	A quitté l'école à 16 ans
-1.43719	A quitté l'école à 17 ans
-1.22751	A quitté l'école à 18 ans
-0.61113	Études collégiales ou universitaires, sans certificat ni diplôme
-0.05921	Certificat/diplôme professionnel
0.45468	Diplôme universitaire
1.16365	Master
1.98437	Doctorat

- Country est le pays de résidence actuel du participant et a l'une des valeurs suivantes :

Value	Meaning
-0.09765	Australie
0.24923	Canada
-0.46841	Nouvelle Zélande
-0.28519	Autres pays
0.21128	Irlande
0.96082	Royaume-Uni
-0.57009	États-Unis

- Ethnicity est l'ethnicité du participant et a l'une des valeurs suivantes :

Value	Meaning
-0.09765	Asiatique
0.24923	Noir
-0.46841	Mixte-Noir/Asiatique
-0.28519	Mixte-Blanc/Asiatique
0.21128	Mixte-Blanc/Noir
0.96082	Autres
-0.57009	Blancs

- Nscore est le Neuroticisme (ou névrosisme) : il caractérise une tendance persistante à l'expérience des émotions négatives. 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 est l'Extraversion : Attitude, comportement d'un individu qui montre une grande facilité à établir des contacts avec ceux qui l'entourent, qui exprime aisément ses sentiments.
- Oscore est l'ouverture à l'expérience : Il s'agit de la tendance d'une personne à s'ouvrir aux expériences, quelles que soient la nature de celles-ci.
- Ascore est l'agréabilité : est un trait de personnalité qui se manifeste dans des caractéristiques comportementales individuelles qui sont perçues comme gentilles, sympathiques, coopératives, chaleureuses et attentionnées.

- Cscore est la conscienciosité : C'est un trait qui consiste à être approfondi, prudent et vigilant et qui implique un désir de bien faire une tâche.
- Impulsive est l'impulsivité mesurée par le BIS-11 (Barrat Impulsivness Scale) : trait de personnalité caractérisé par un comportement direct adopté par un individu sans que celui-ci ne pense aux conséquences de ses actes.
- SS est la recherche de sensation mesurée par ImpSS : elle se définit par le besoin d'expériences et de sensations variées, complexes, pouvant conduire le sujet à s'engager dans des conduites de désinhibition, des activités physiques et sociales risquées.

III.Description des targets

i. Alcohol est la classe de consommation d'alcool.

ii. Amphet est la classe de consommation d'amphétamines.

iii. Amyl est la classe de la consommation de nitrite d'amyle.

iv. Benzos est la classe de consommation de benzodiazépines.

v. Caff est la classe de consommation de caféine.

vi. Cannabis est la classe de la consommation de cannabis.

vii. Choc est la classe de la consommation de chocolat.

viii. Coke est la classe de consommation de cocaïne.

ix. Le crack est une catégorie de consommation de crack.

x. Ecstasy est la classe de consommation d'ecstasy.

xi. Heroin est la classe de consommation d'héroïne.

xii. Ketamine est la classe de consommation de kétamine.

xiii. Legalh est la classe de consommation de legal highs.

xiv. LSD est la classe de la consommation de LSD.

xv. Meth est la classe de consommation de méthadone.

xvi. Mushrooms est la classe de la consommation de champignons magiques.

xvii. Nicotine est la classe de la consommation de nicotine.

xviii. Semer est la classe de consommation de la drogue fictive Semeron.

xix. VSA est la classe de consommation de substances volatiles.

Approche du problème

Exploratory Data Analysis

Objectif : comprendre au maximum les données dont on dispose pour définir une stratégie de modélisation.

1. Analyse de la forme:
   - Identification de la target : plusieurs choix s'offrent à nous. Il faut qu'on trouve une façon de répartir les consommations de drogues, car il y a 7 cas possibles ce qui fait beaucoup.
   - Première approche du dataset
   - Nombre des lignes et de colonnes
   - Identification des valeurs manquantes
   - Types de variables

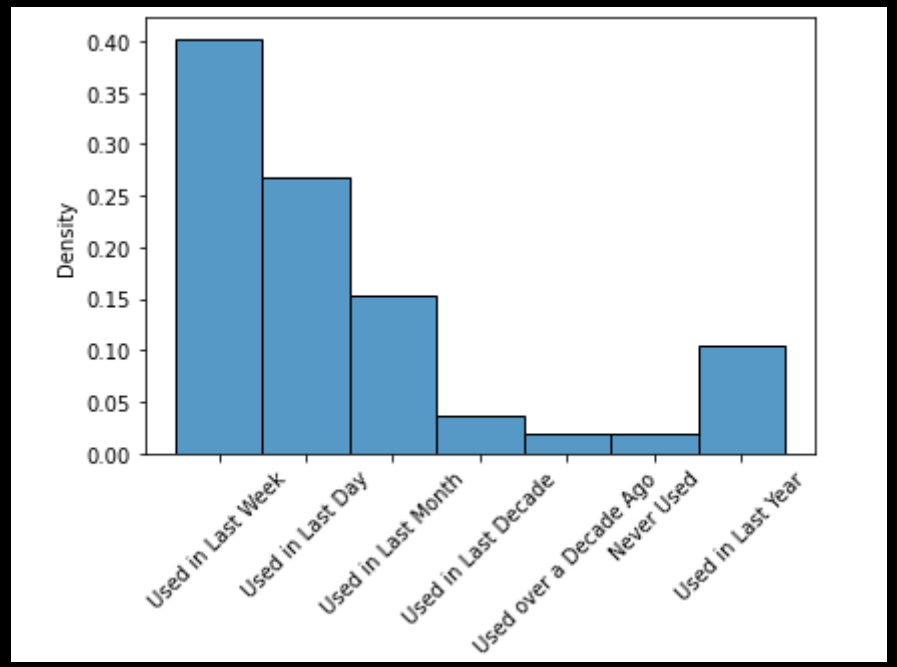
Object	Float	Int
19	12	1

2. Analyse du fond

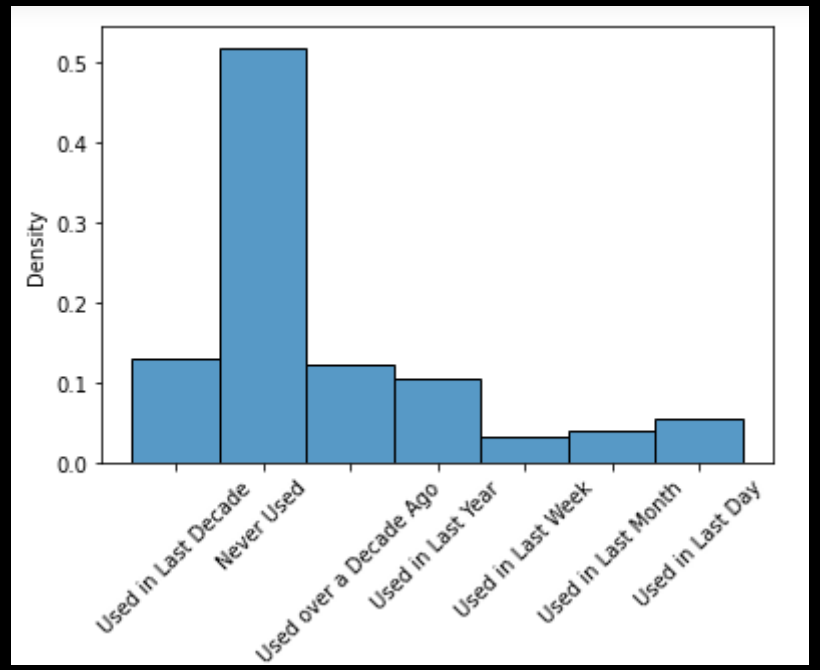
- Visualisation et compréhension des targets (histogramme/boxplot)
- Compréhension des différentes variables (recherche)
- Visualisation des relations : features/targets
- Relation variable / variable

Visualisation de la target pour toutes les drogues

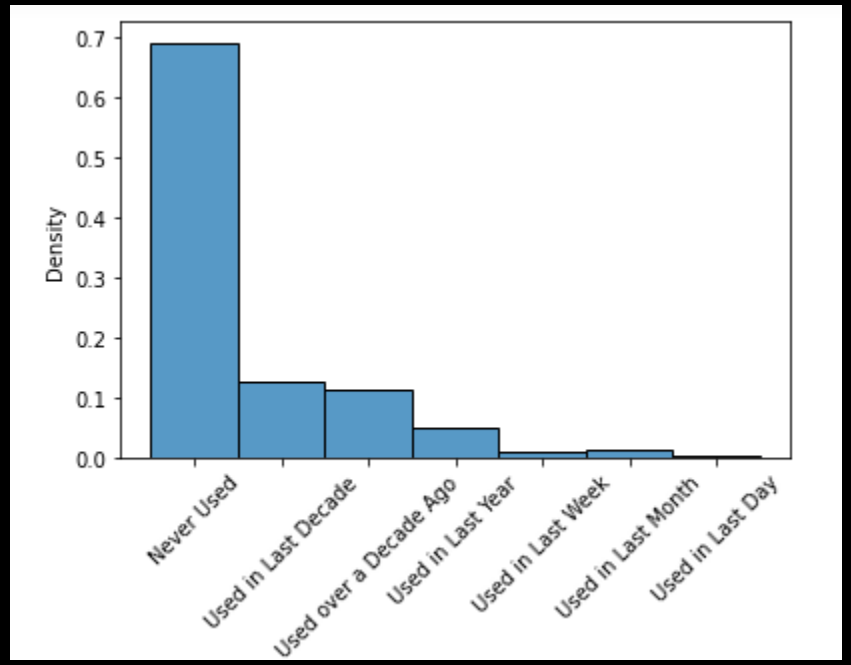
- Alcool :



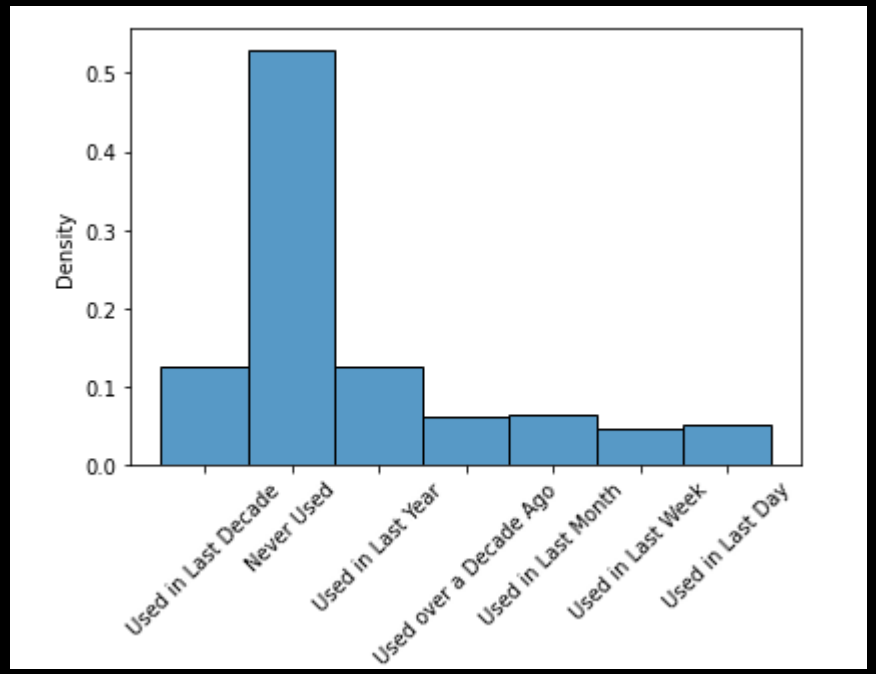
- Amphétamines



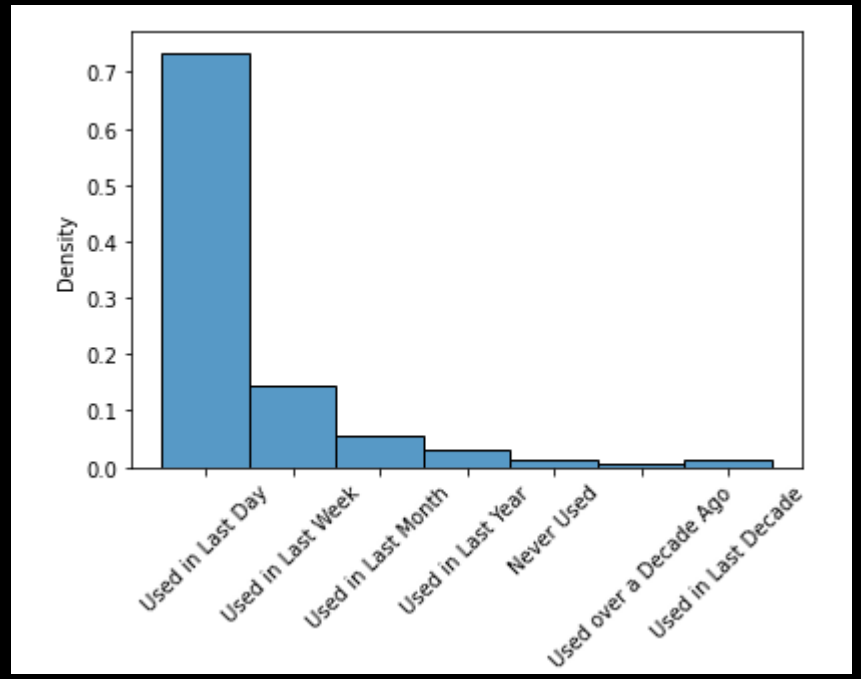
- Amyl


Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Never Used
Used in Last Decade
Used over a Decade Ago
Used in Last Year
Used in Last Week
Used in Last Month
Used in Last Day

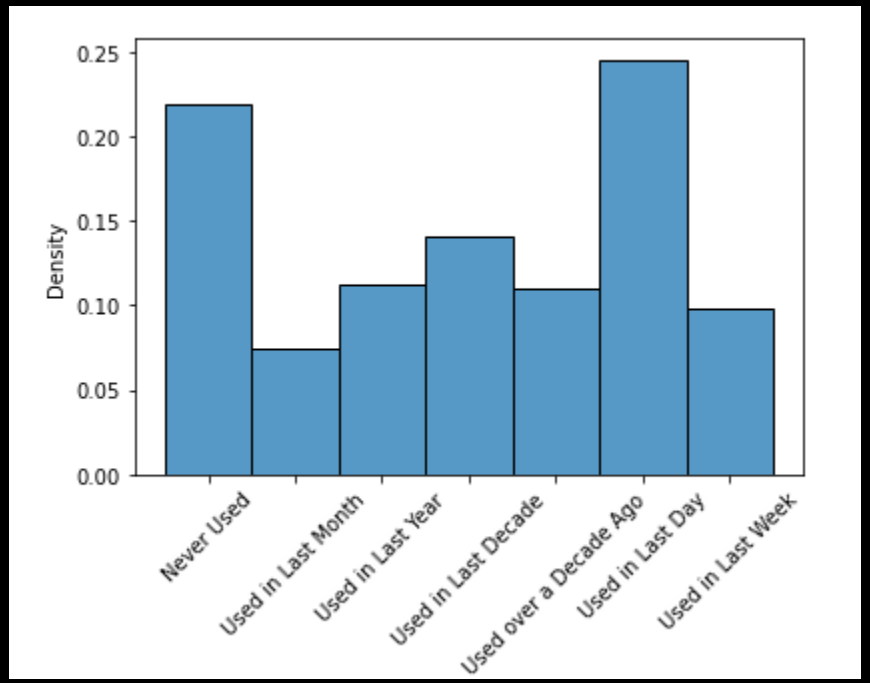
- Benzo


Density
0.0
0.1
0.2
0.3
0.4
0.5
Used in Last Decade
Never Used
Used in Last Year
Used over a Decade Ago
Used in Last Month
Used in Last Week
Used in Last Day

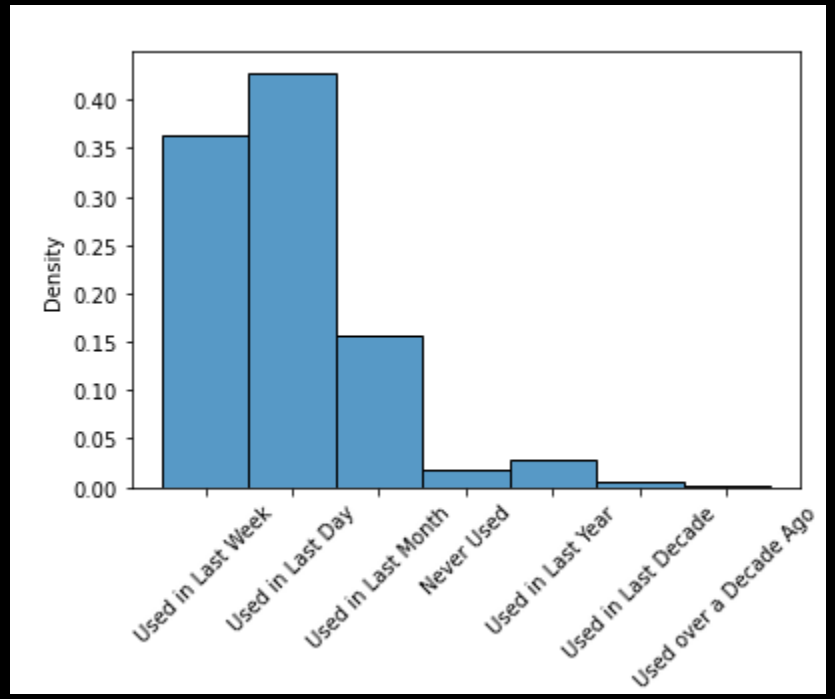
- Caffeine



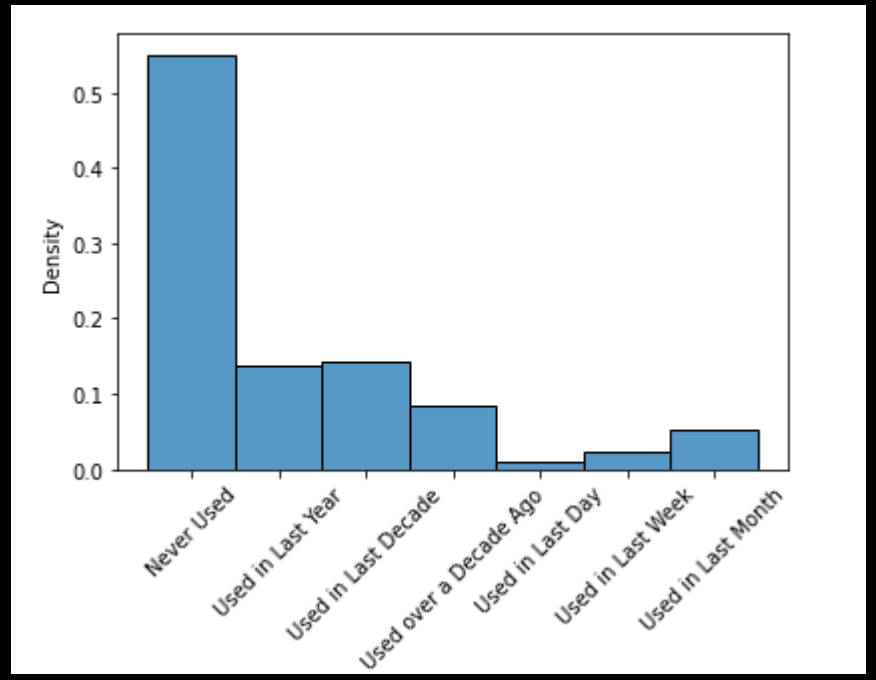
- Cannabis



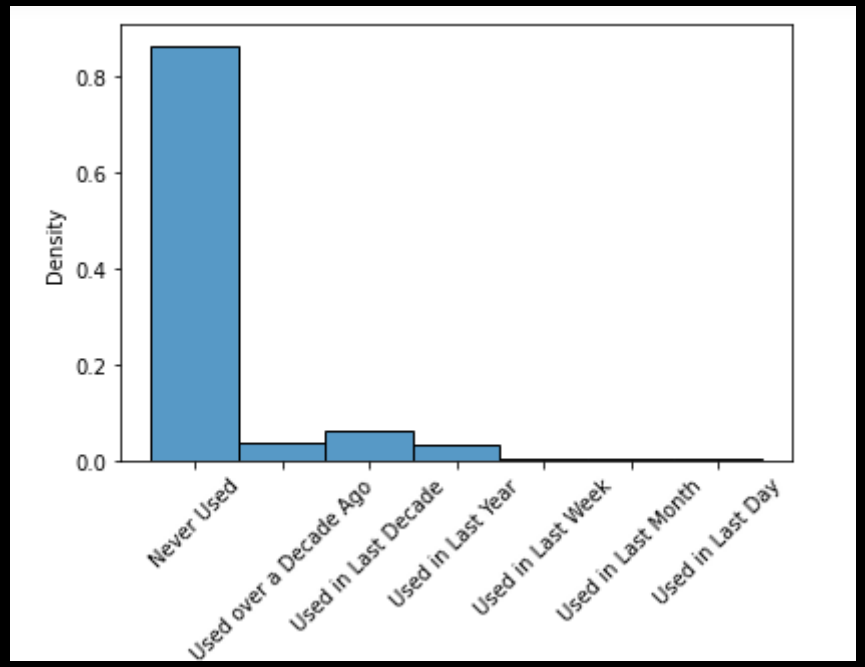
- Chocolate



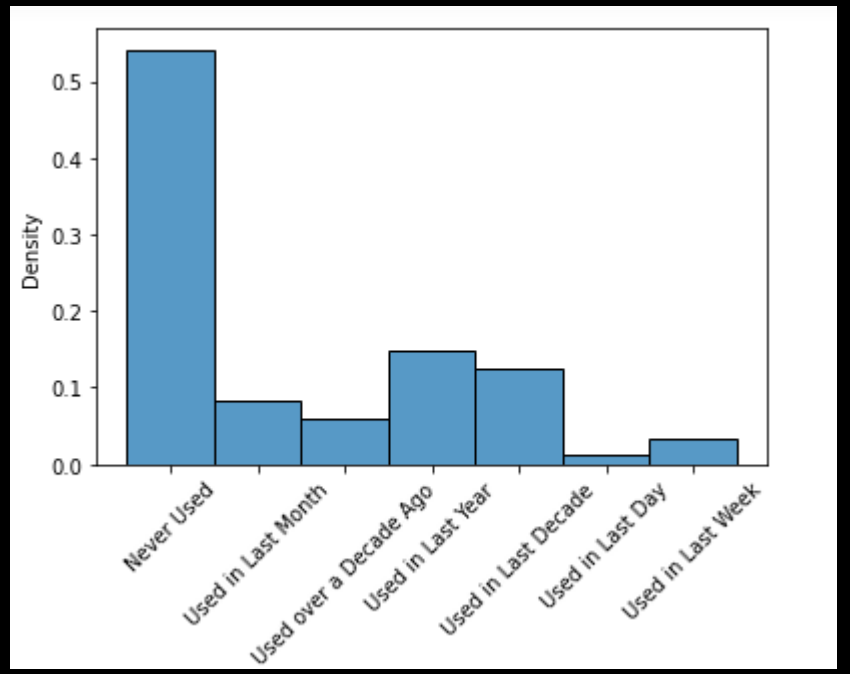
- Cocaine



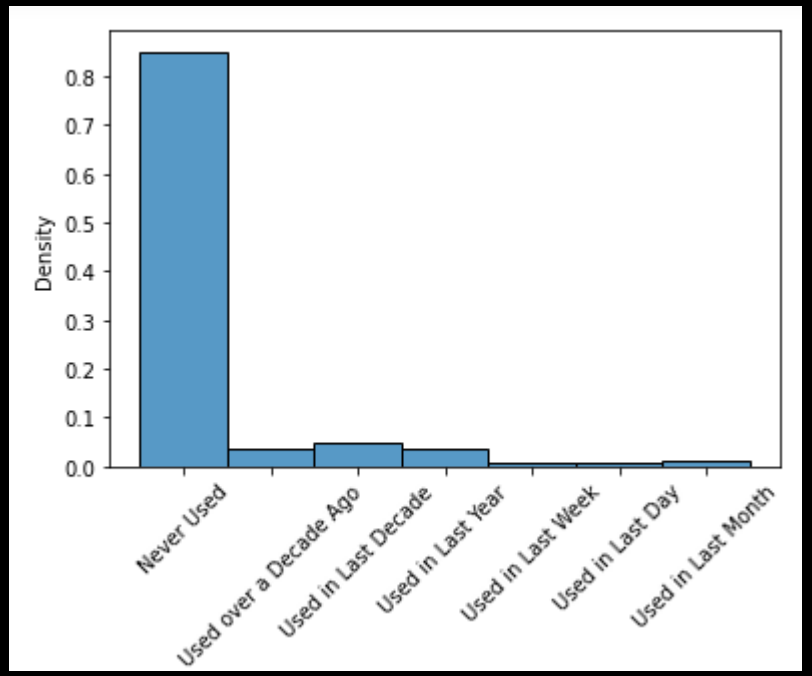
- Crack



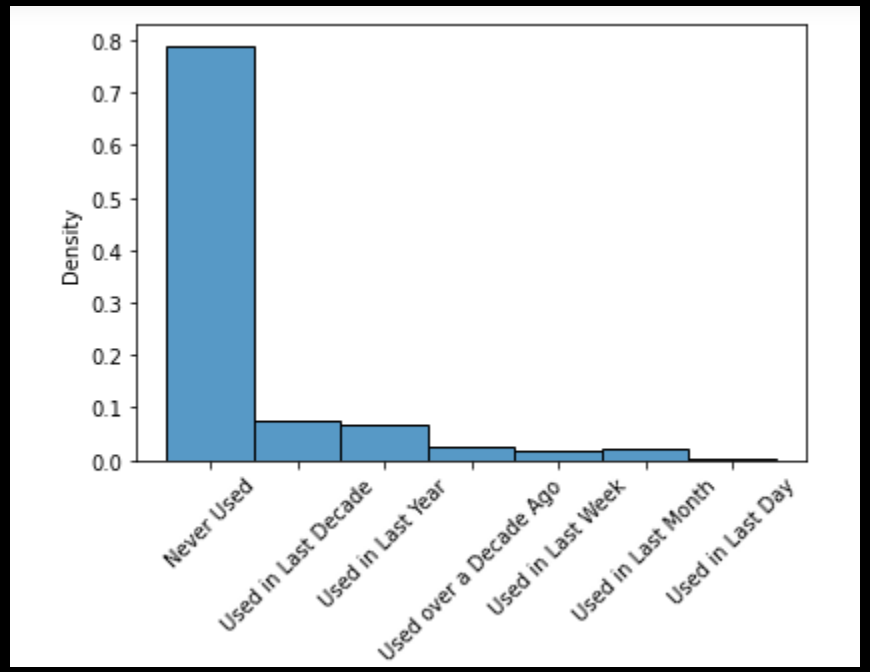
- Ecstasy



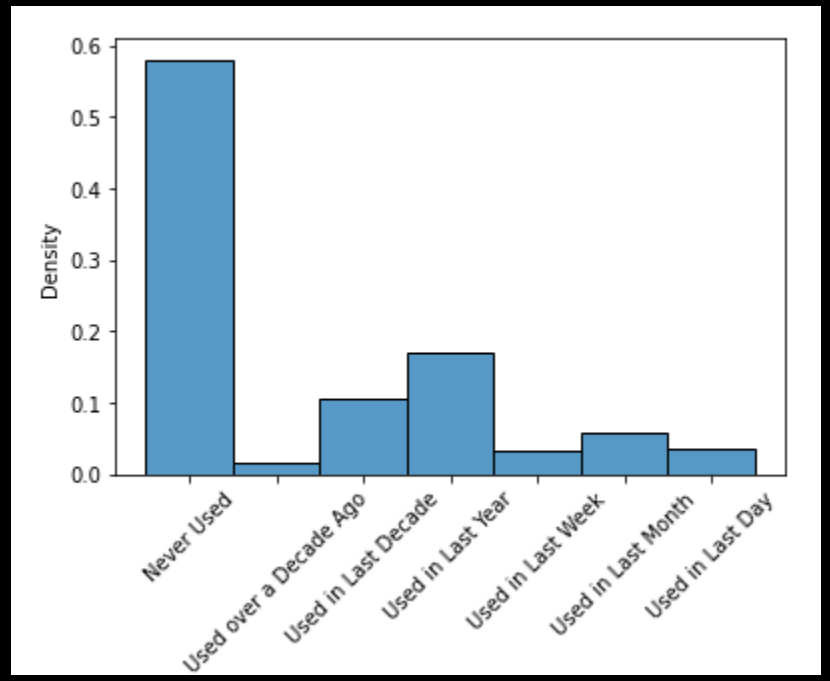
- Heroin



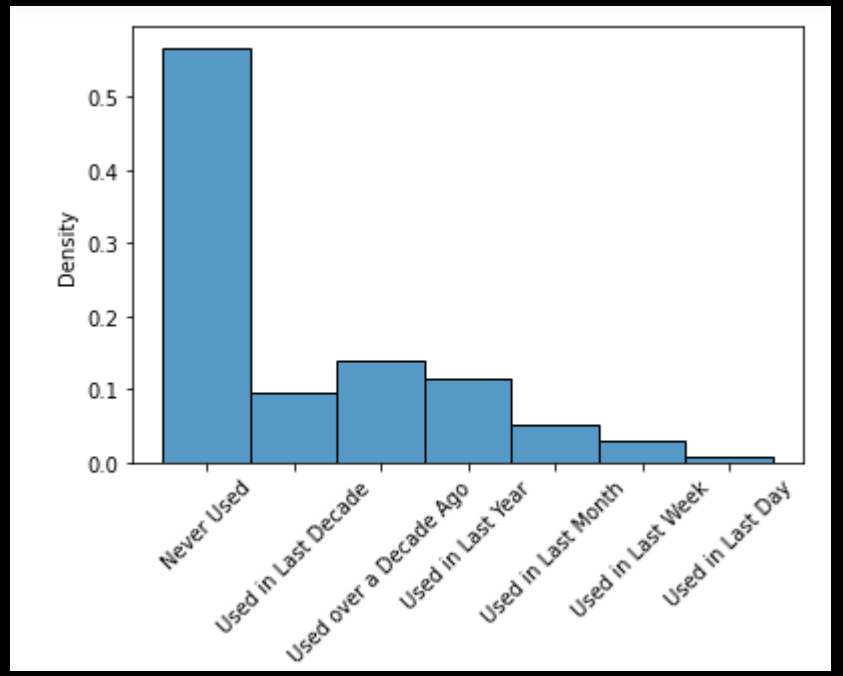
- Ketamine



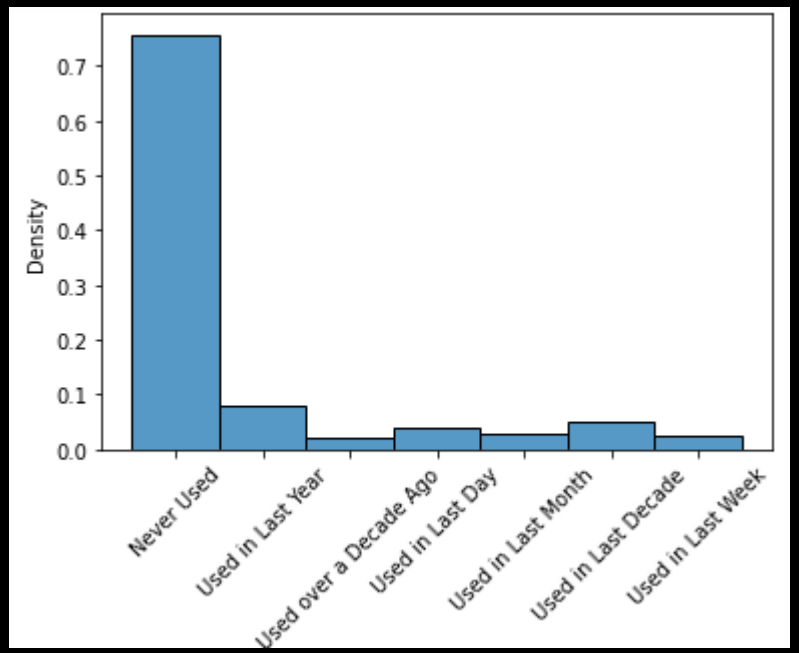
- Legal highs



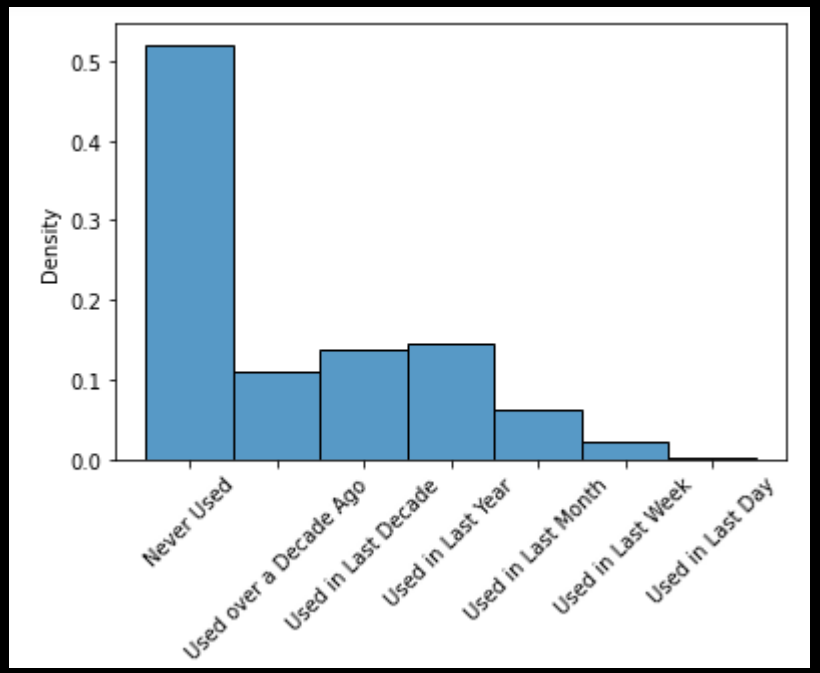
- LSD



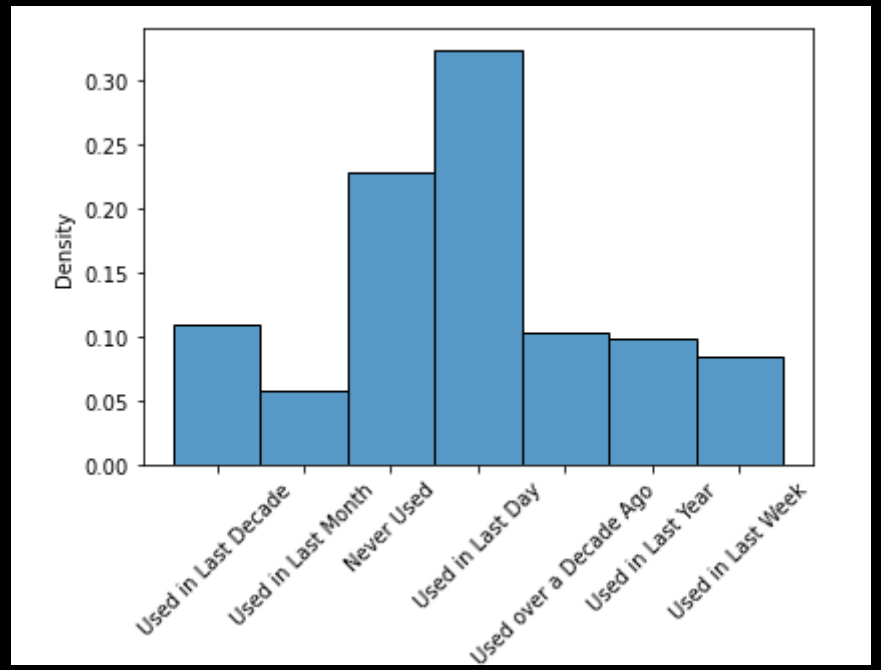
- Methadone



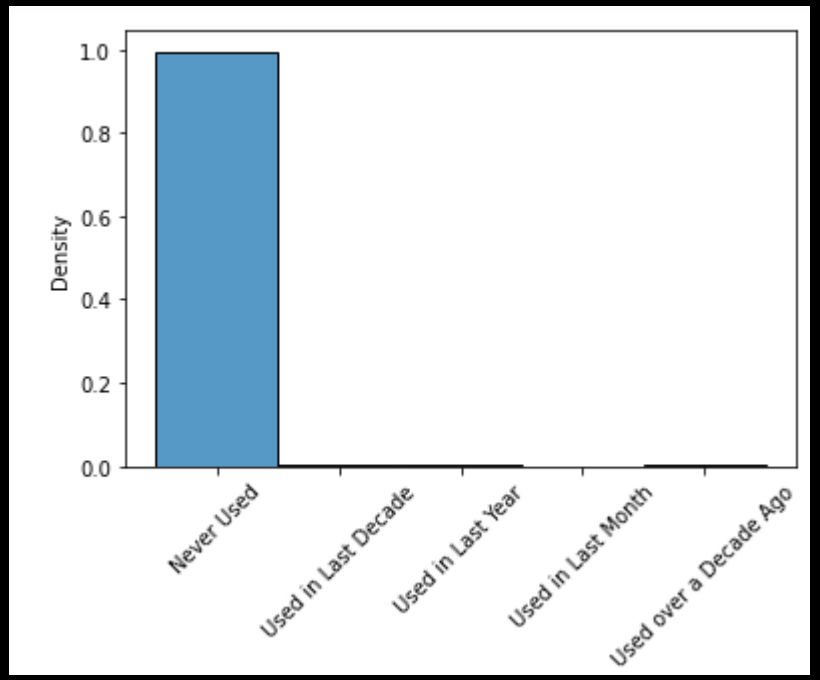
- Magic mushrooms



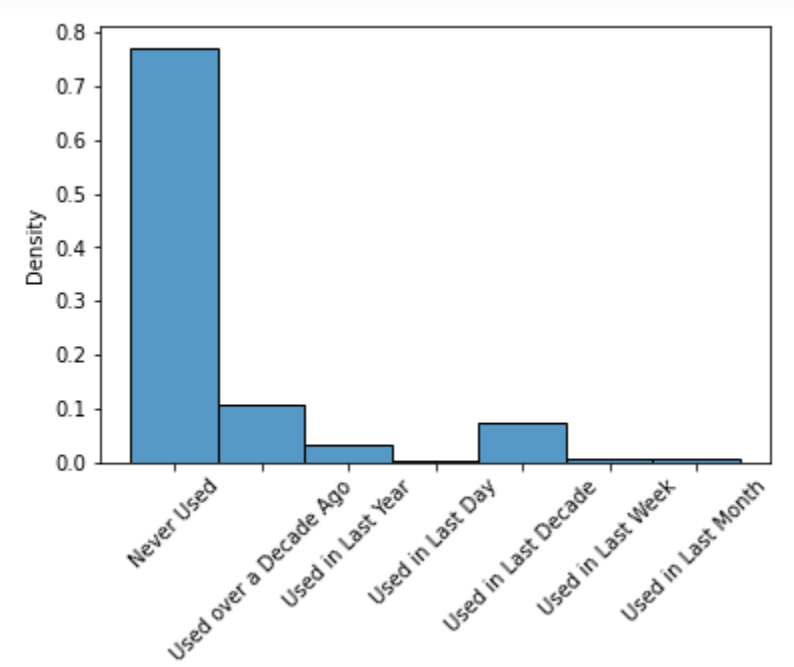
- Nicotine



- Semer



- VSA



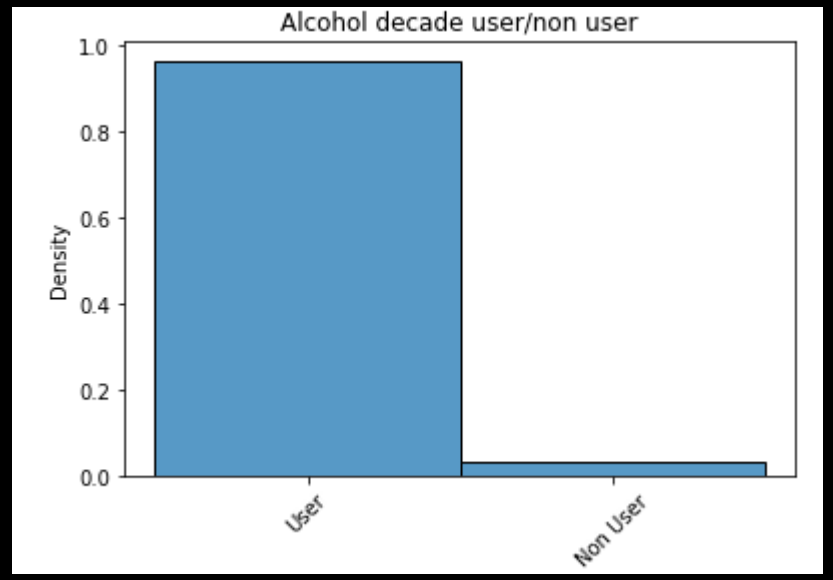
- Globalement, la répartition des classes est peu équilibrée, il s'agirait de changer la répartition des targets en regroupant les 7 catégories en 2 (binarisation).
- Par exemple, une classification basée sur la consommation par décennies, ou par mois. Tout dépend de ce qu'on veut faire, ce qu'on veut étudier. Il est vrai qu'une étude qui détermine si on va être un consommateur régulier ou pas, se baserait sur une répartition par mois de la consommation ({never used/used over a decade ago/used in last decade/used in last year} {used in last month/used in last week/used in last day})

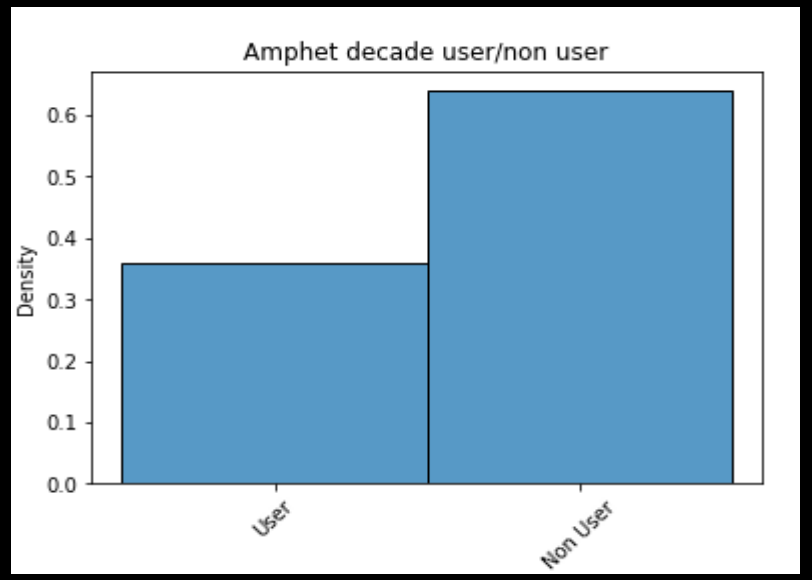
Pour pouvoir comparer l’équilibre des classes que proposent les différents types de binarisation on pourrait utiliser le critère d’entropie de Shannon. Néanmoins nous allons nous baser sur la binarisation basée sur les décades car elle paraît globalement plus équilibrée que notamment la binarisation par mois.

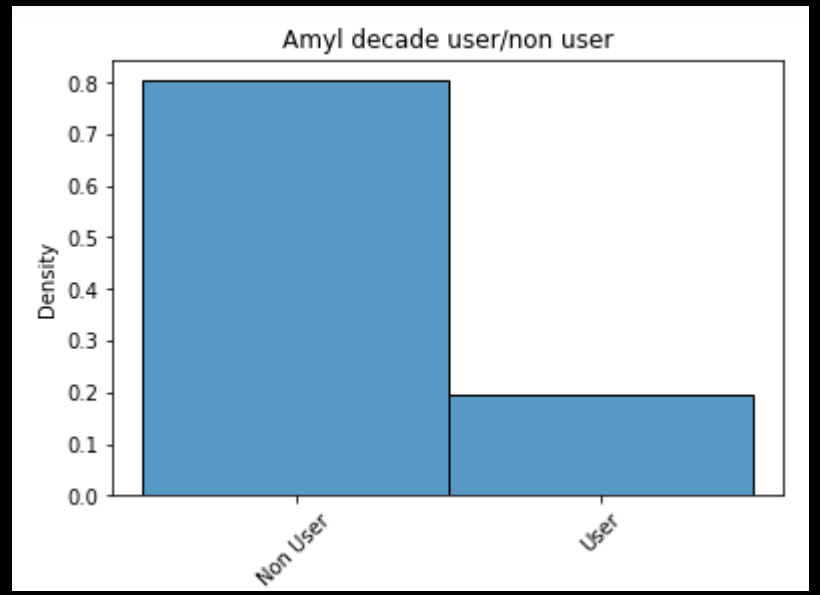
Nous utiliserons, dans la suite de notre travail, cette binarisation.

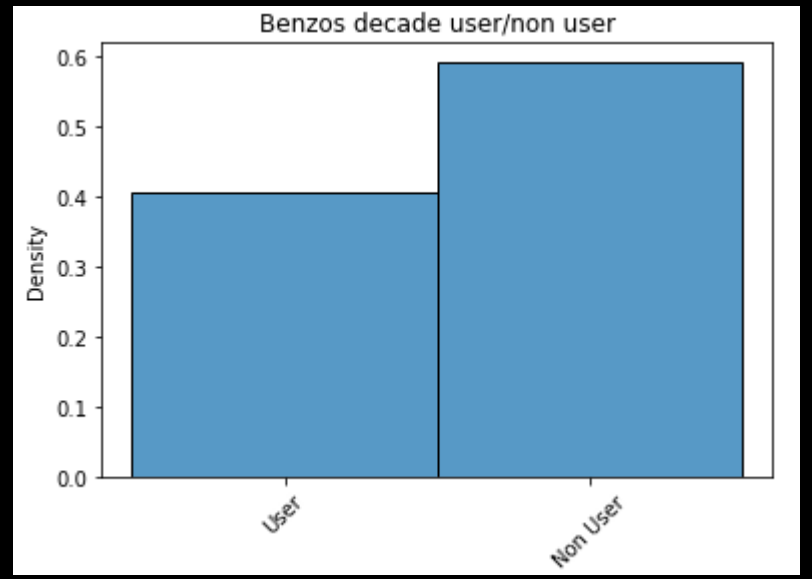
Nous pouvons voir (sur la diapositive suivante) les visualisations de la target pour toutes les drogues en se basant sur la binarisation "decade user/non user".

Visualisation de la target pour toutes les drogues(basée sur la décennie)

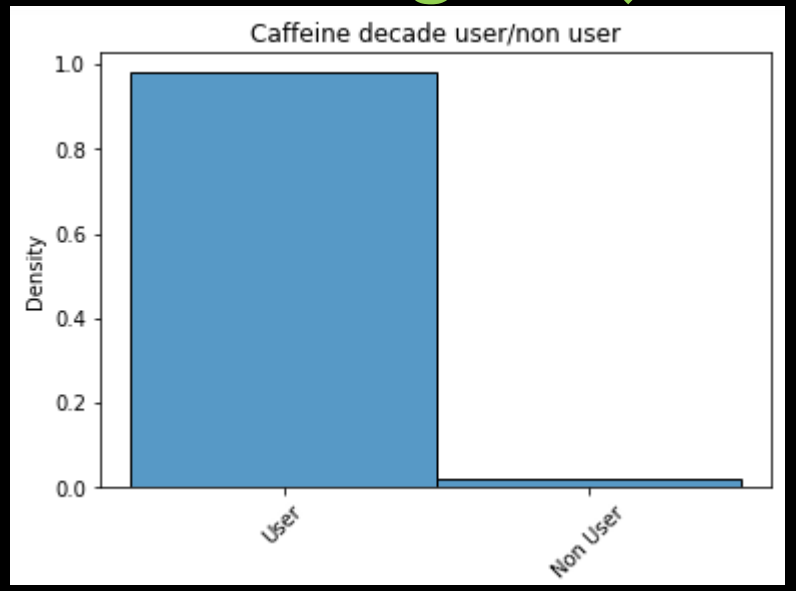


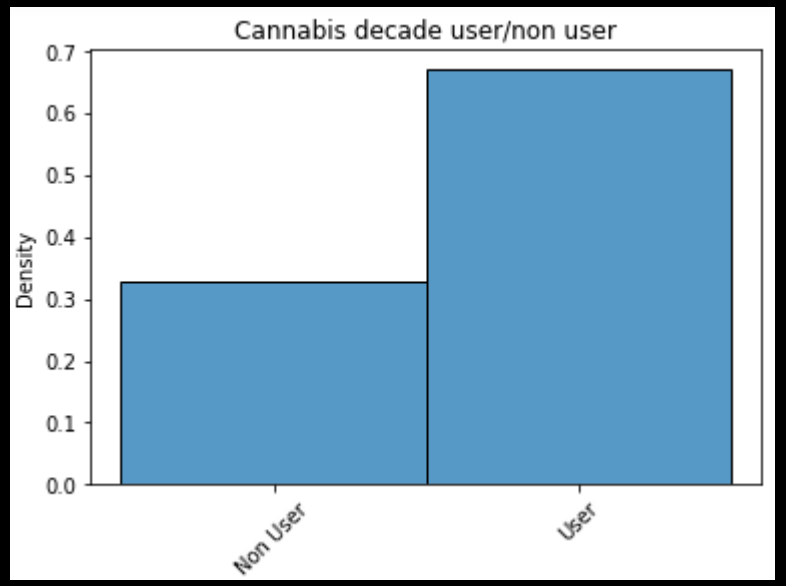


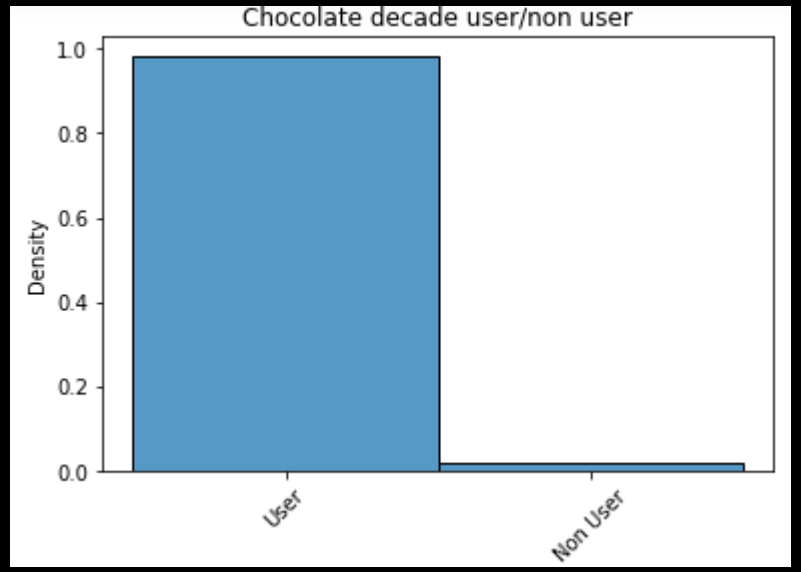


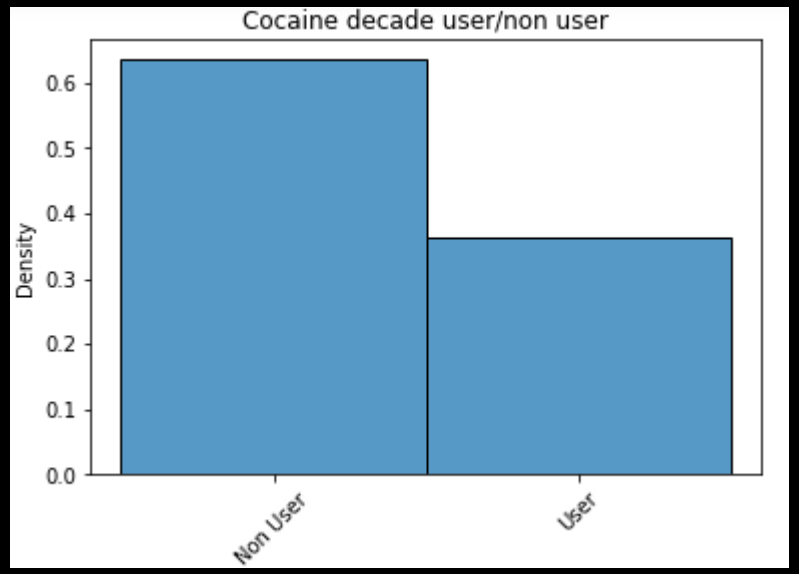


Visualisation de la target pour toutes les drogues(basée sur la décennie)

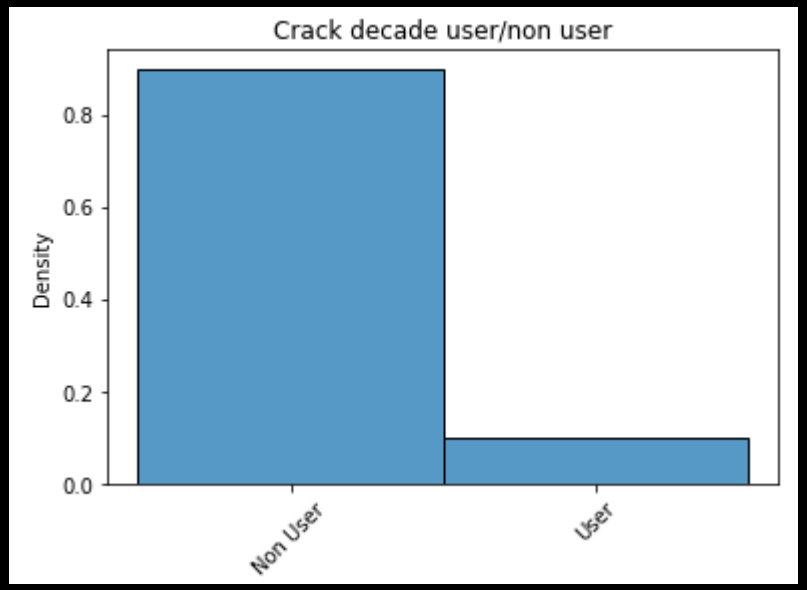


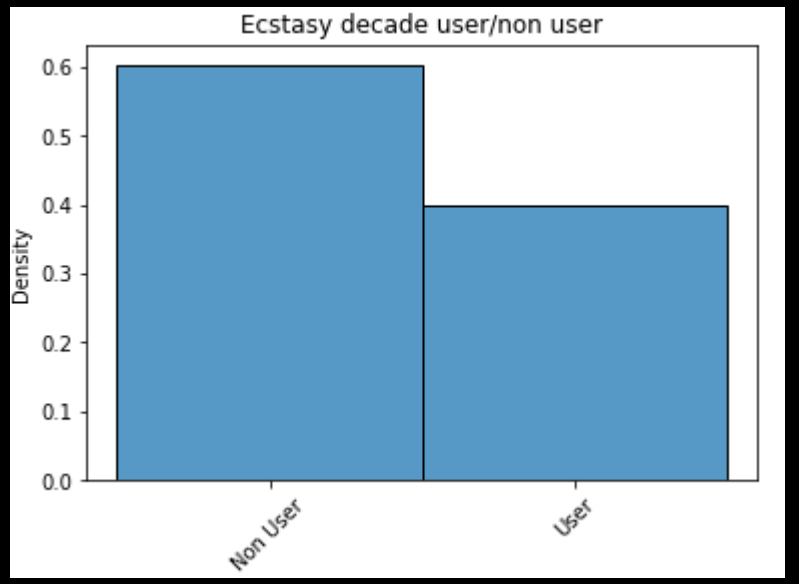


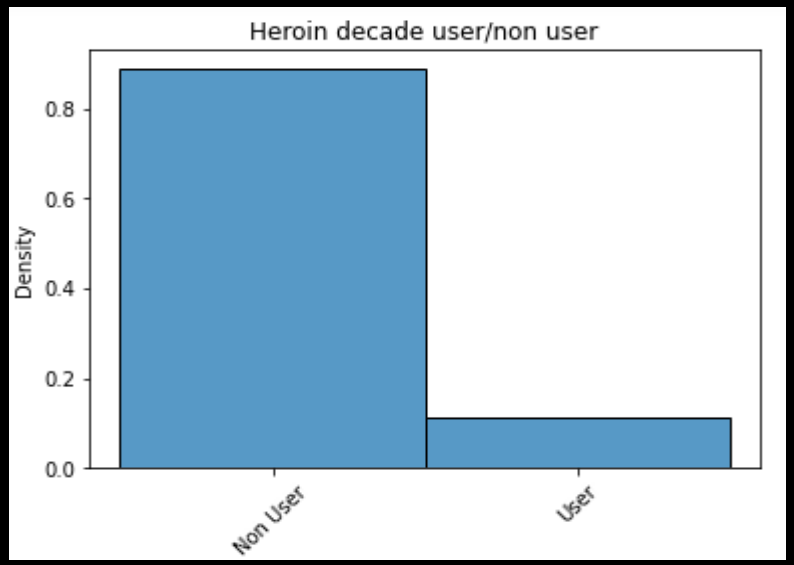


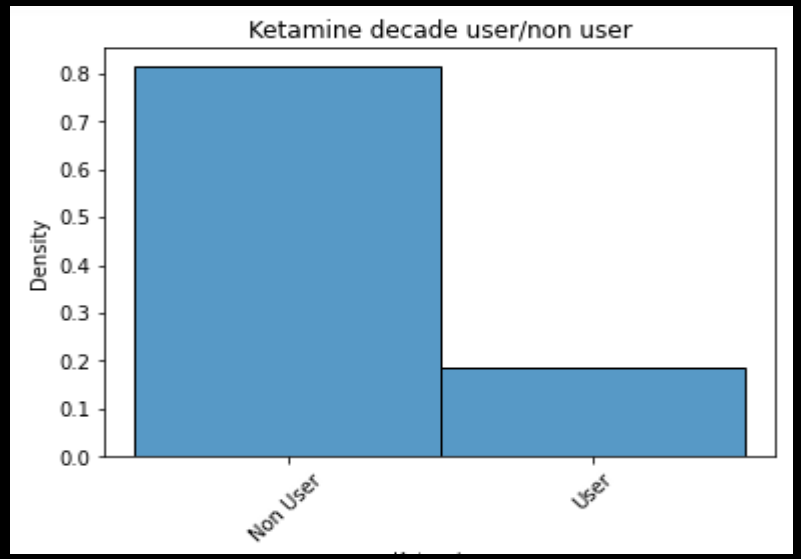


Visualisation de la target pour toutes les drogues(basée sur la décennie)

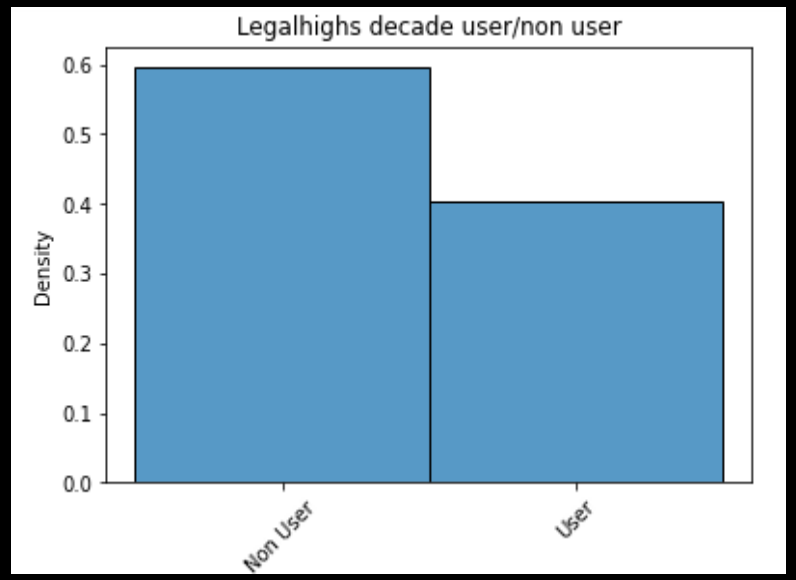


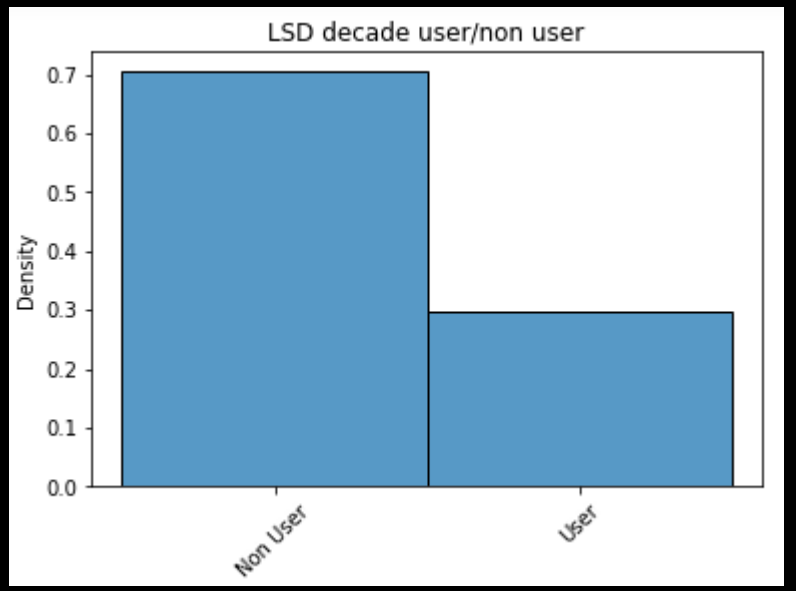


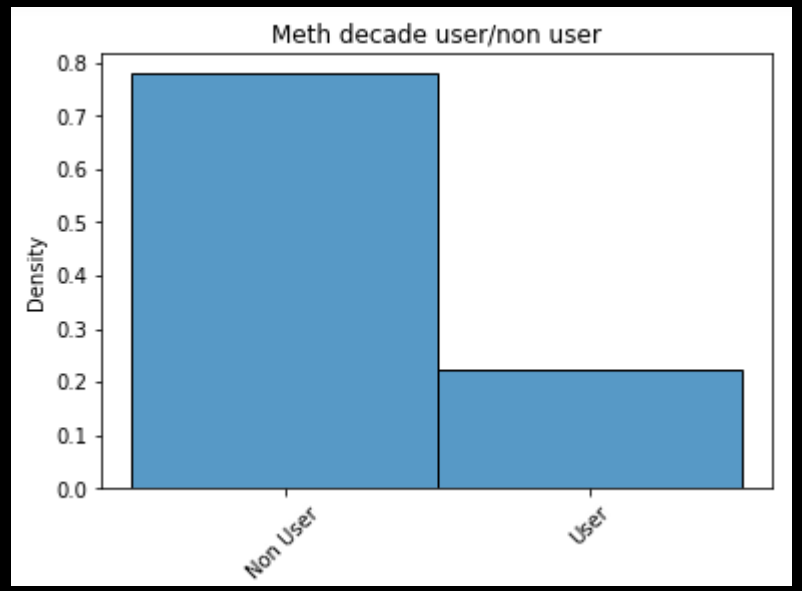


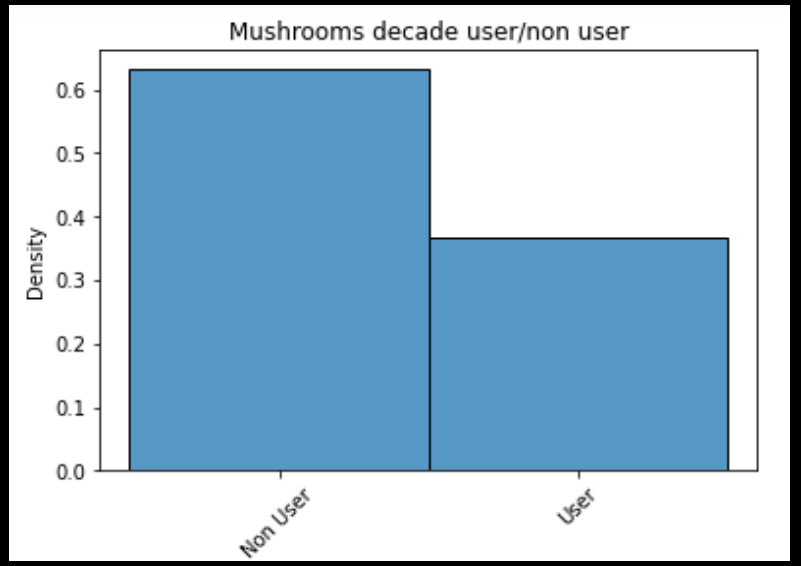


Visualisation de la target pour toutes les drogues(basée sur la décennie)

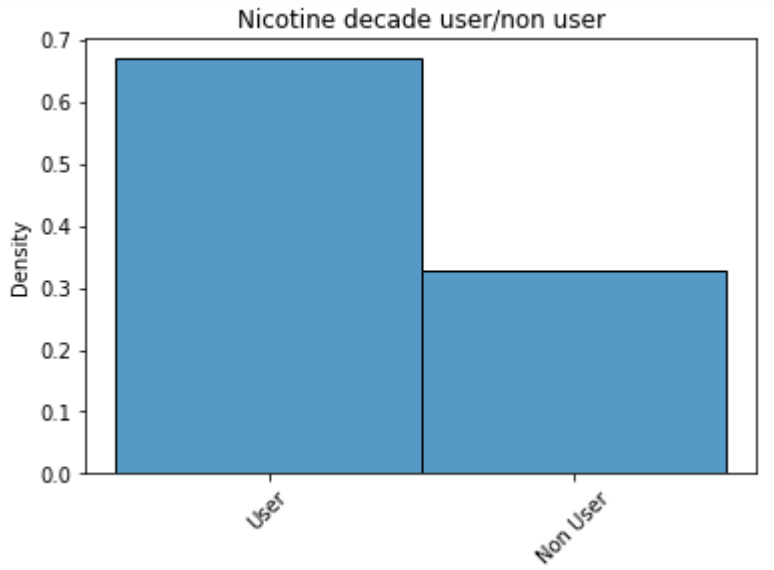


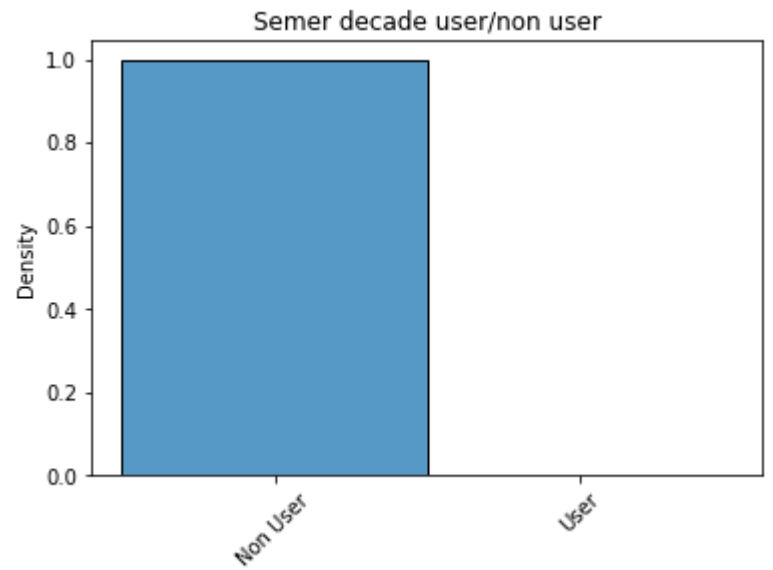


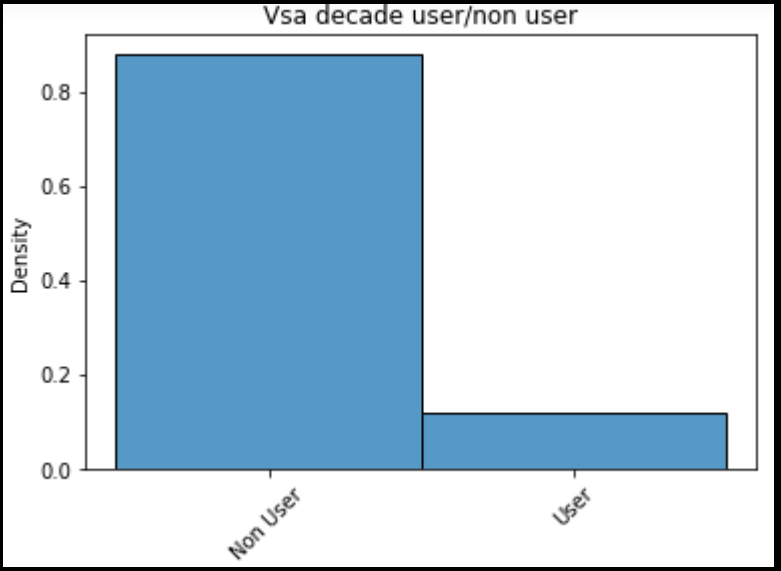




Visualisation de la target pour toutes les drogues(basée sur la décennie)

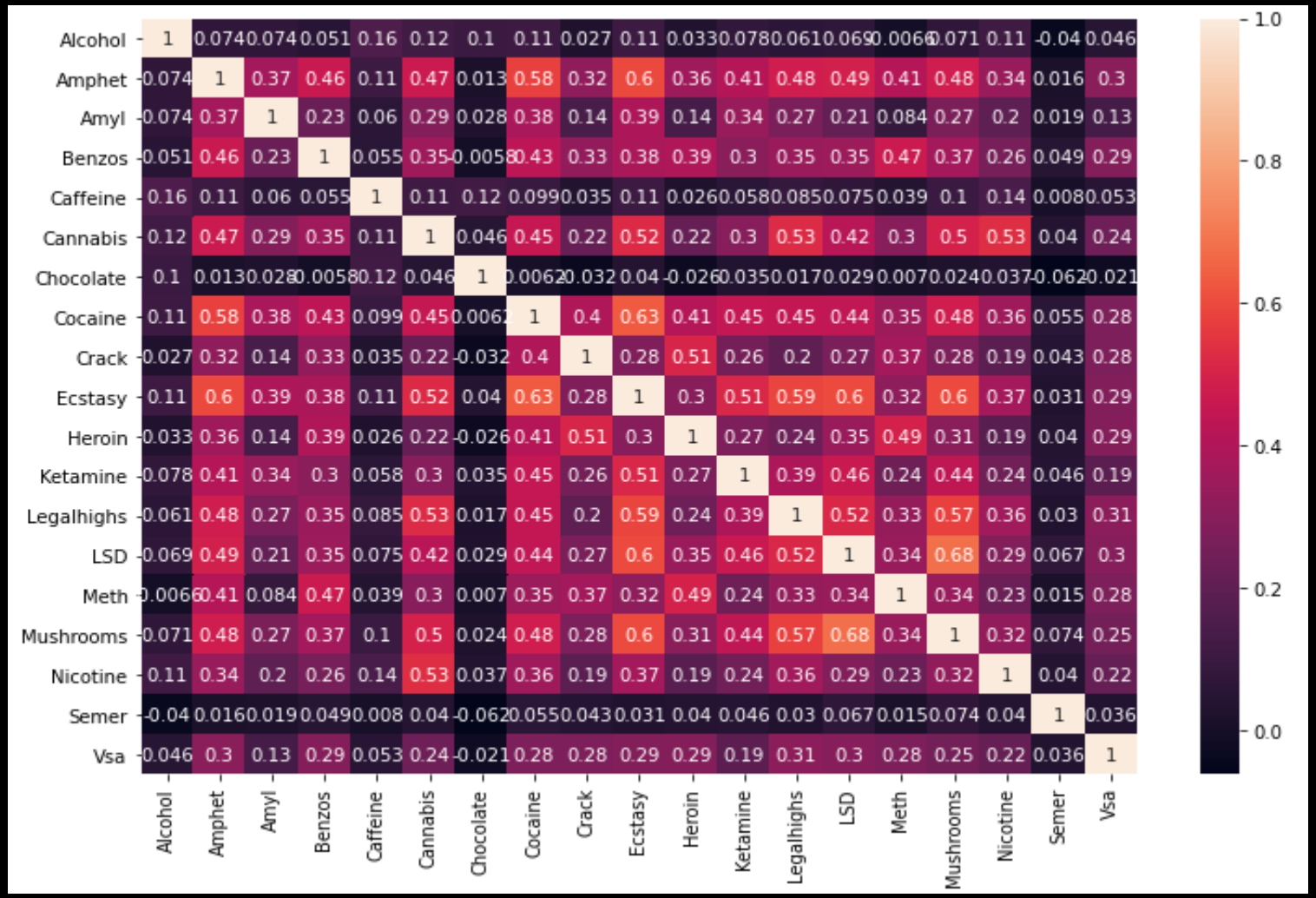






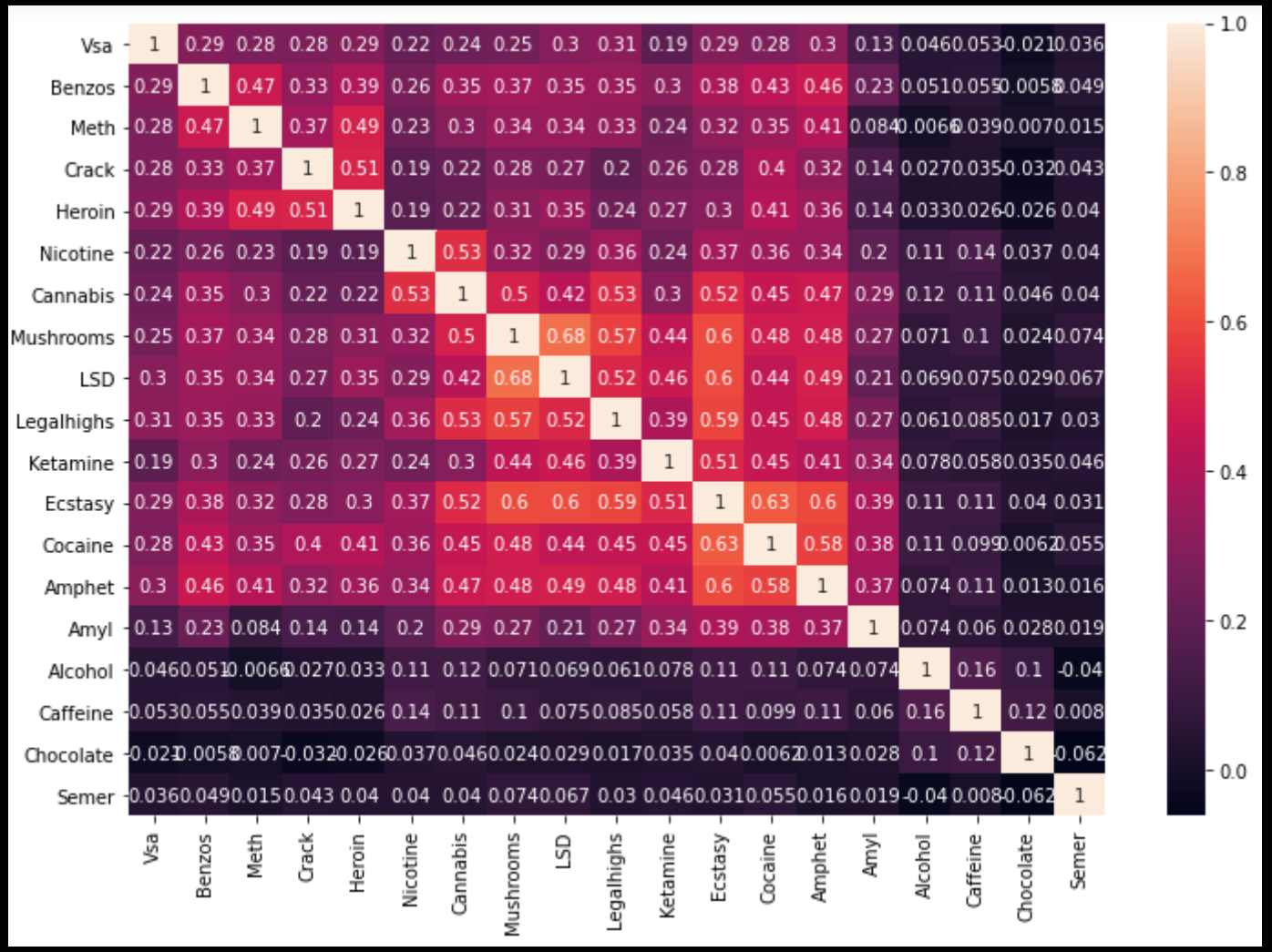
Visualisation des corrélations entre les drogues

Maintenant voyons, grâce à la matrice de corrélation, comment les drogues sont corrélées entre elles:



Clustering

Objectif : faire une étude généralisée sur les drogues les plus corrélées entre elles afin de réduire le nombre des targets



Les pléiades

- Nous avons lu un article qui a démontré qu'il existe trois groupes de drogues dont la consommation est fortement corrélée, en d'autres termes, ces drogues sont très souvent consommées ensemble. L'idée de fusionner des attributs corrélés en "modules" est populaire en biologie. Les modules sont appelés les "pléiades de corrélation".
  - Pléiade de l'héroïne (HeroinPl) constituée de : héroïne, crack, cocaïne et méthadone.
  - Pléiade de l'ecstasy (EcstasyPl) constituée de : ecstasy, amphétamines, cannabis, cocaïne, kétamine, LSD, magic mushrooms, legal highs.
  - Pléiade de la benzo (BenzoPl) constituée de : benzo, méthadone, cocaïne, amphétamines.
- Les clusters que nous avons trouvés sont certes intéressants mais ne correspondent pas aux pléiades que nous attendions au vu de la documentation que nous avons consultée.

Après avoir créé ces pléiades on peut créer des dataframes pour chaque pléiade et ensuite voir la répartition de user vs non-user.

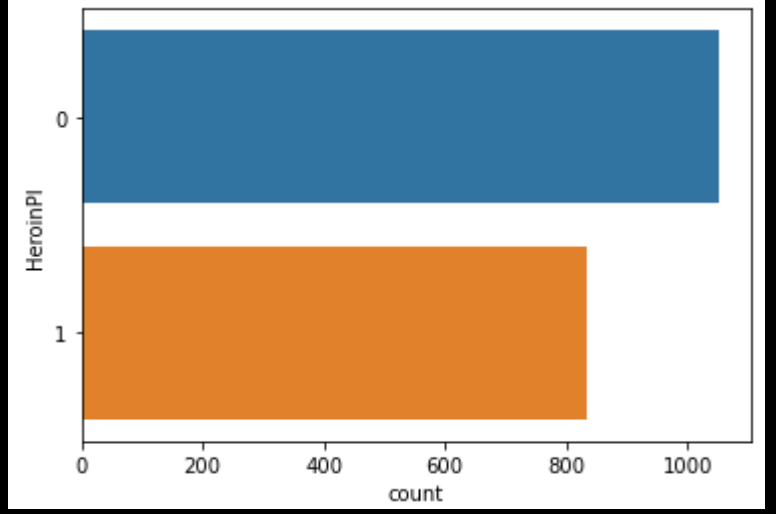
- 1 si un répondant a consommé au moins une des drogues d'une pléiade.
- 0 si il n'en a consommé aucune.

On peut ensuite visualiser la répartition des classes.

Visualisation des pléiades

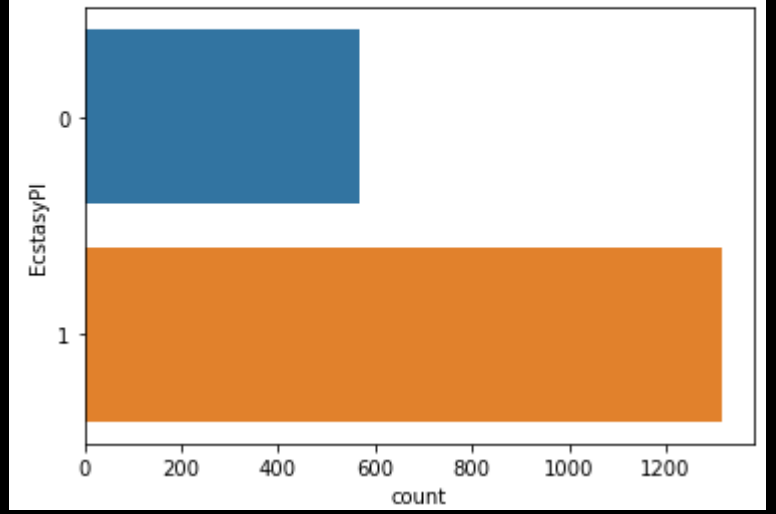
HeroinPl

```
0    55.86
1    44.14
Name: HeroinPl, dtype: float64
```



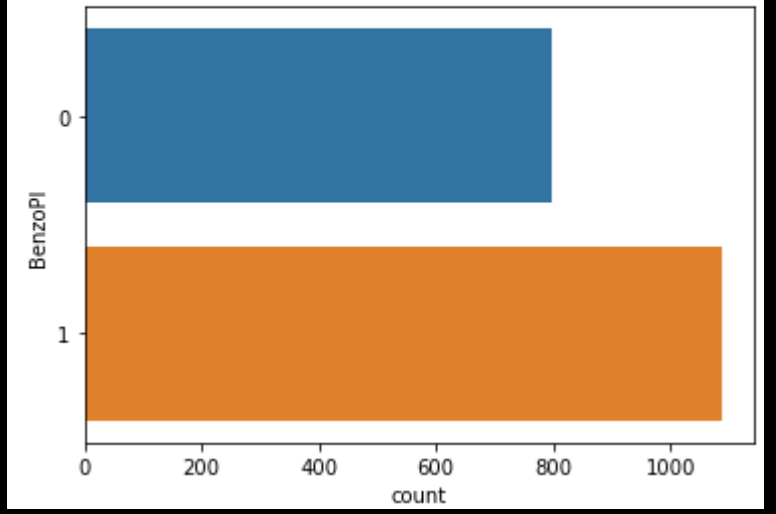
EcstasyPl

```
1    69.87
0    30.13
Name: EcstasyPl, dtype: float64
```



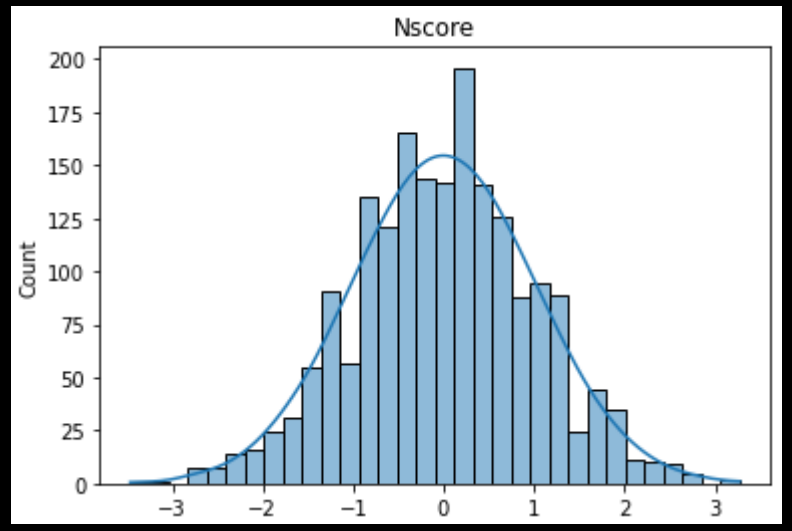
BenzoPl

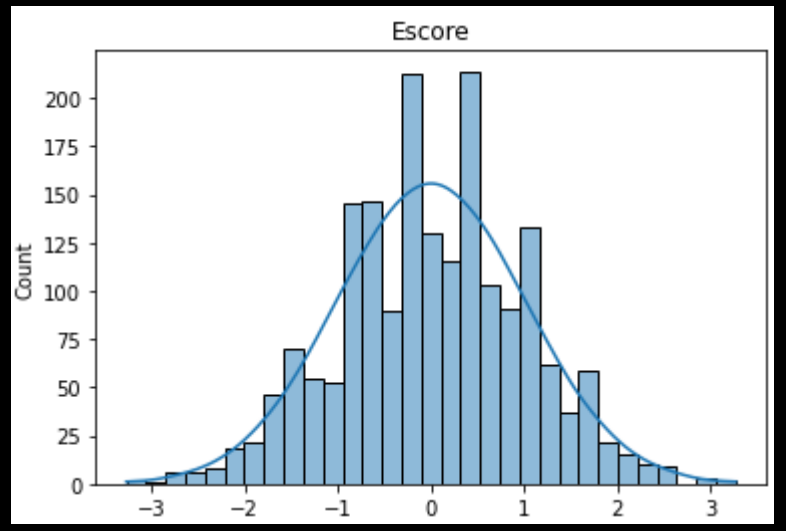
```
1    57.77
0    42.23
Name: BenzoPl, dtype: float64
```

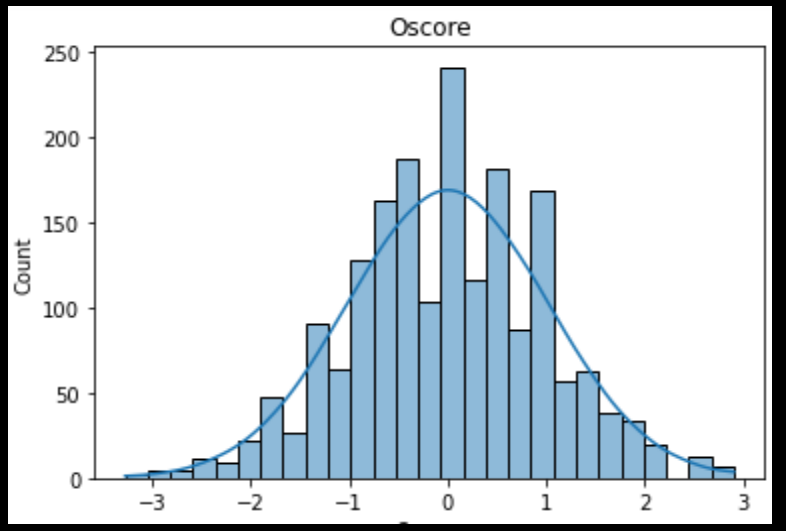


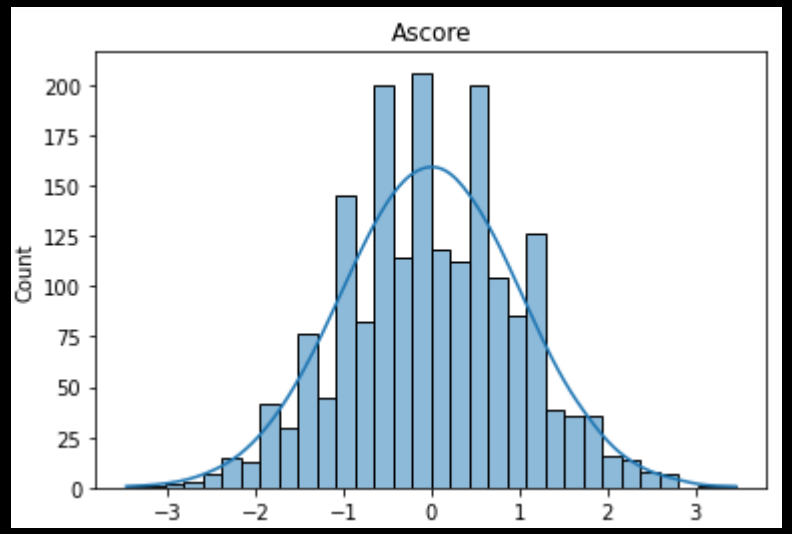
On peut voir que la répartition de classe dans les pléiades est plutôt équilibrée pour l'Héroïne et la Benzo, mais pas pour l'ecstasy (quasiment 70/30). On va utiliser ces pléiades pour visualiser les relations target/feature, car c'est bien plus pratique.

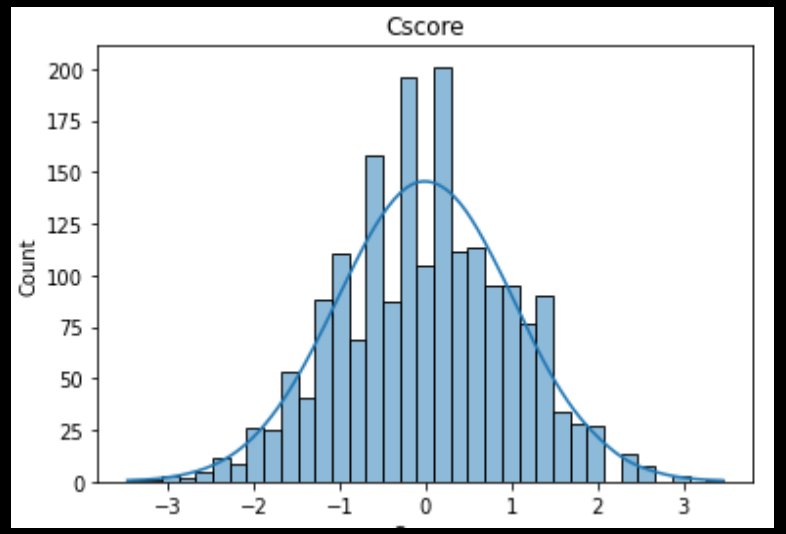
Visualisation de la distribution des features

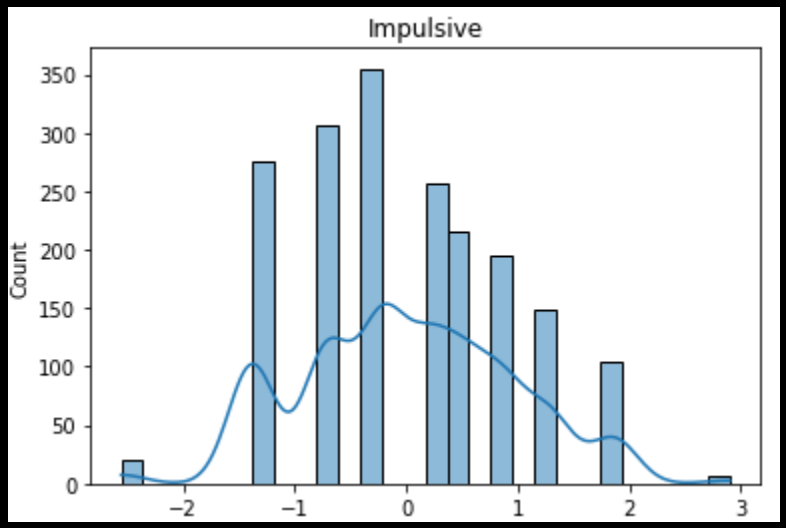




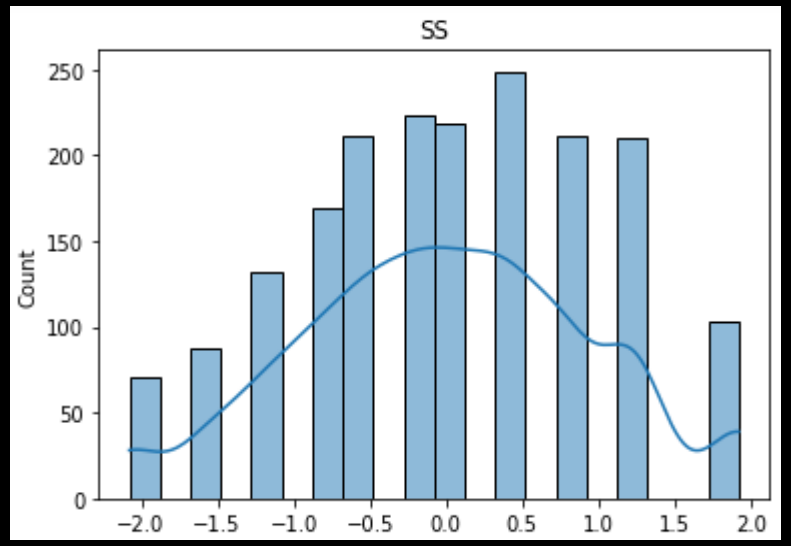


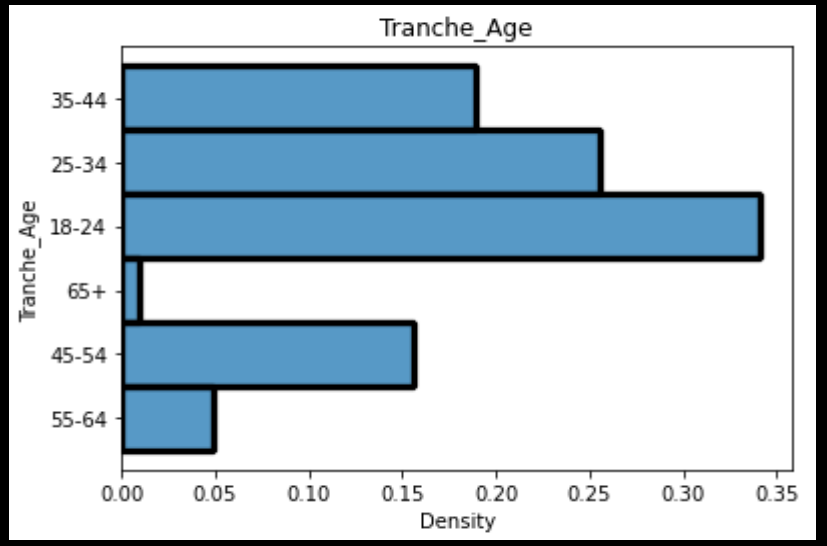


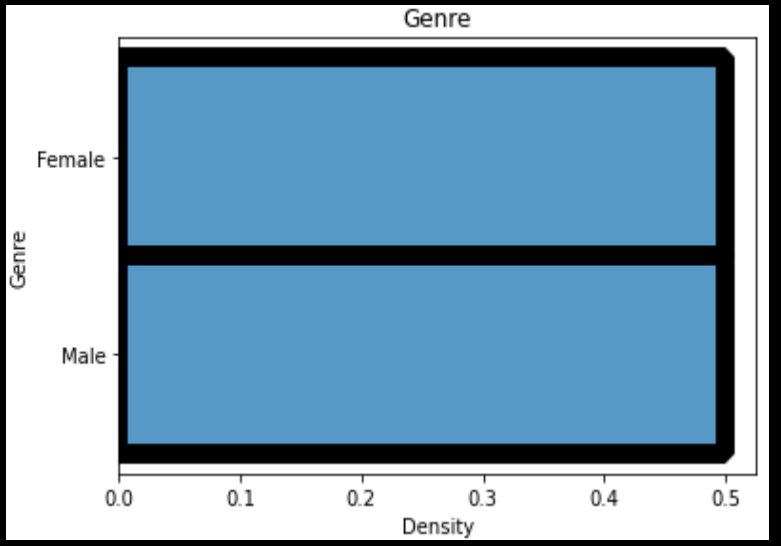


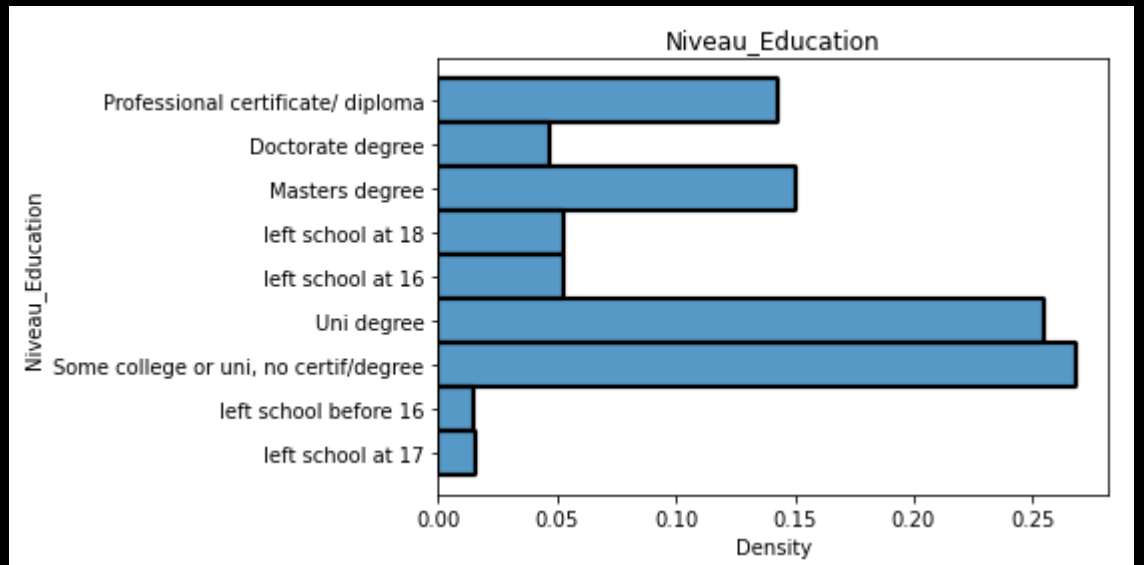


Visualisation de la distribution des features

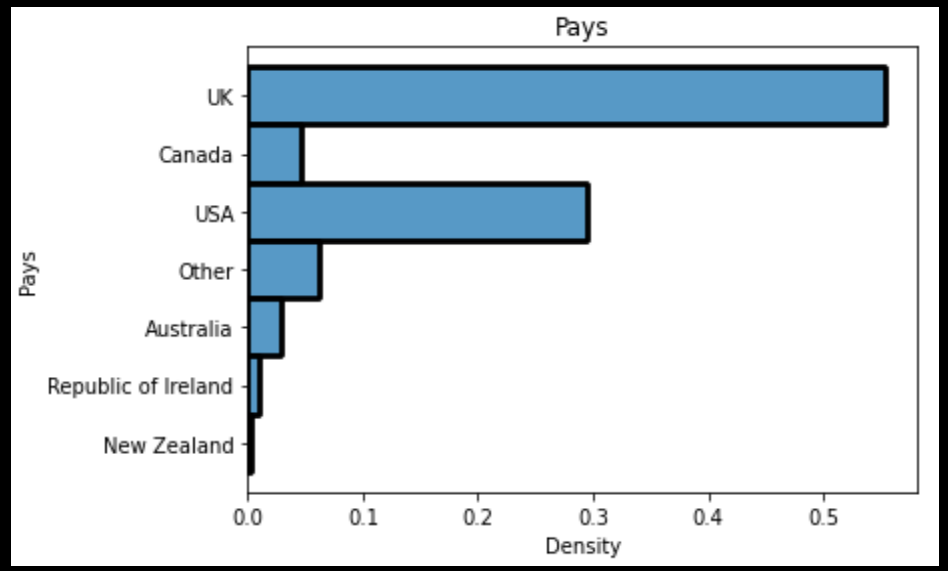


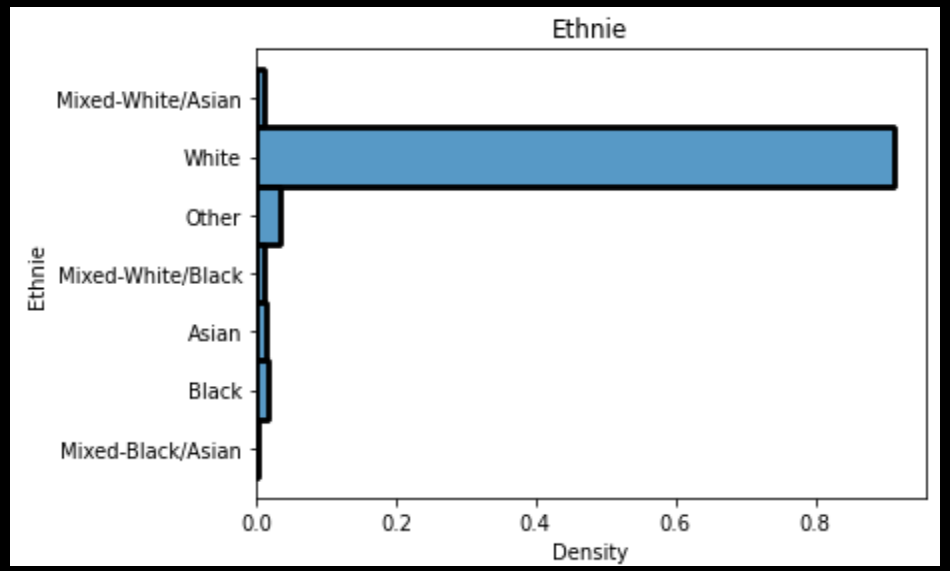






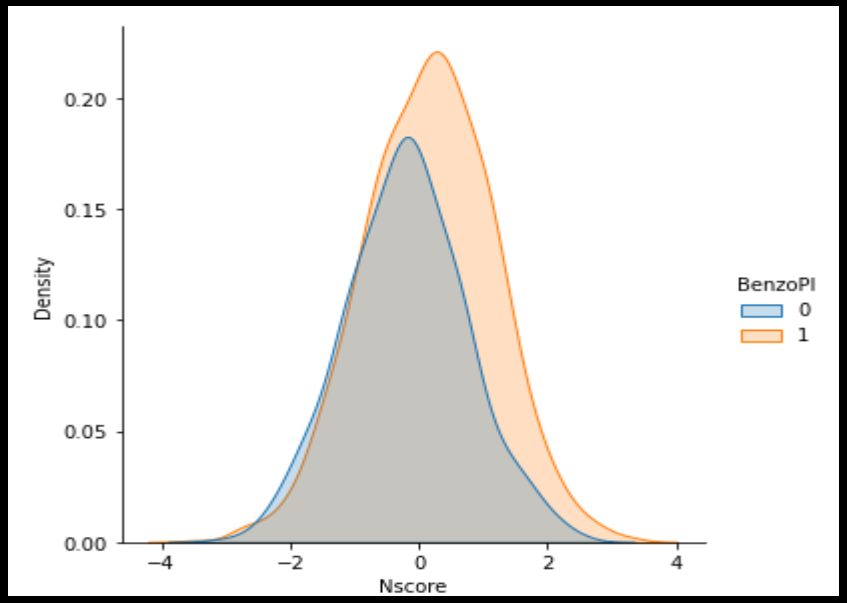
Visualisation de la distribution des features

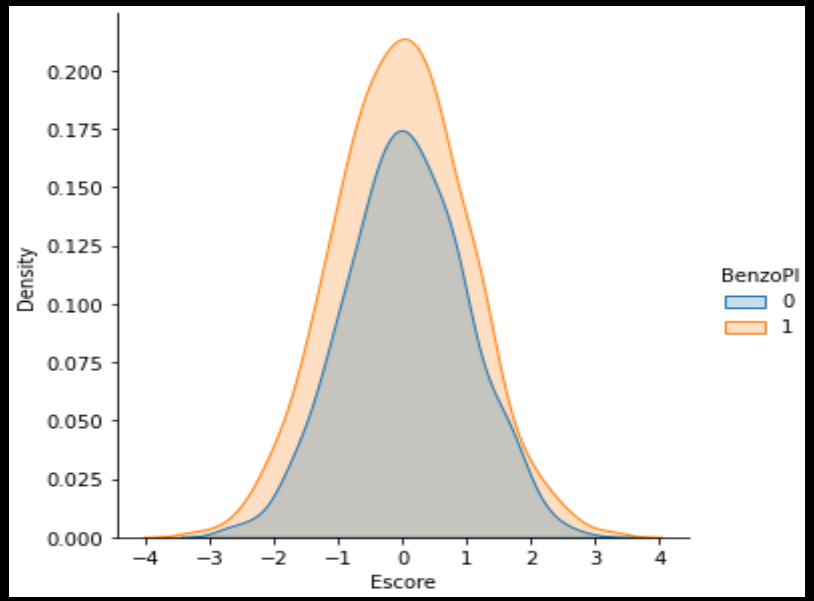


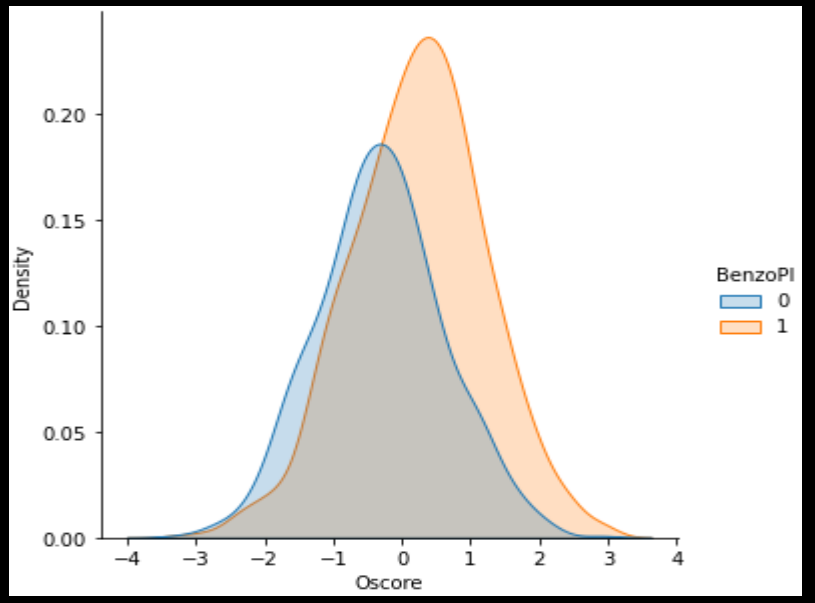


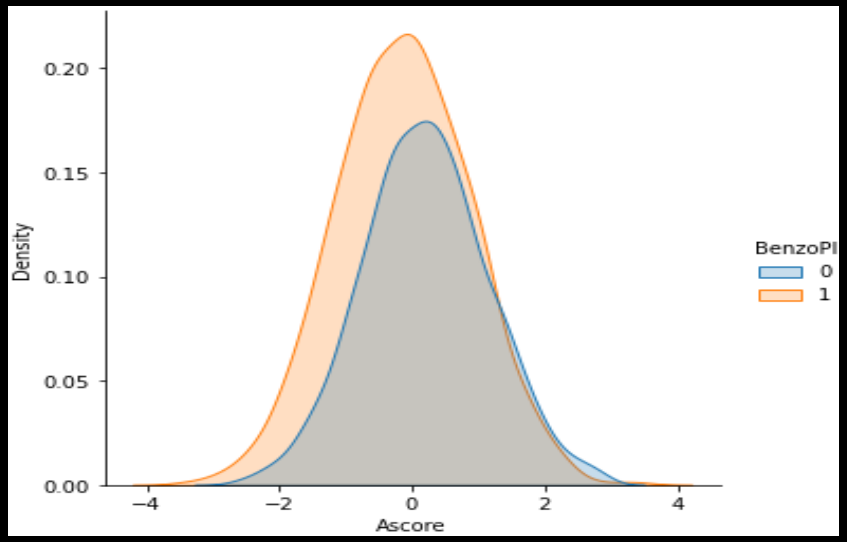
- A priori la répartition des différents scores (N, E, A, O, C) suivent une distribution gaussienne. On ne peut pas en dire autant de SS et Impulsiveness.
- En observant la répartition des classes, on peut voir que les variables ethnie et pays, on l'air d'avoir des variances très faibles. Ethnie est composée à plus de 85% de blanc, et la variable de pays est composée à plus de 80% d'anglais et d'américain. Il parait compliqué de dire que ces catégories peuvent être représentatives dans un modèles prédictifs, avec uniquement 1885 valeurs et une répartition aussi peu équilibrée. On pense ne pas utiliser ces 2 variables, avec les données fournies par le dataset.

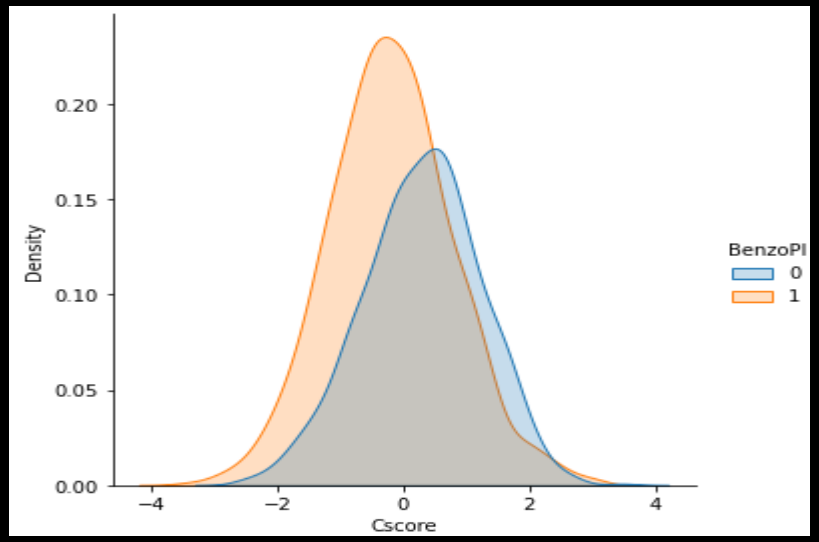
Visualisation des dépendances des features/targets (BenzoPl)

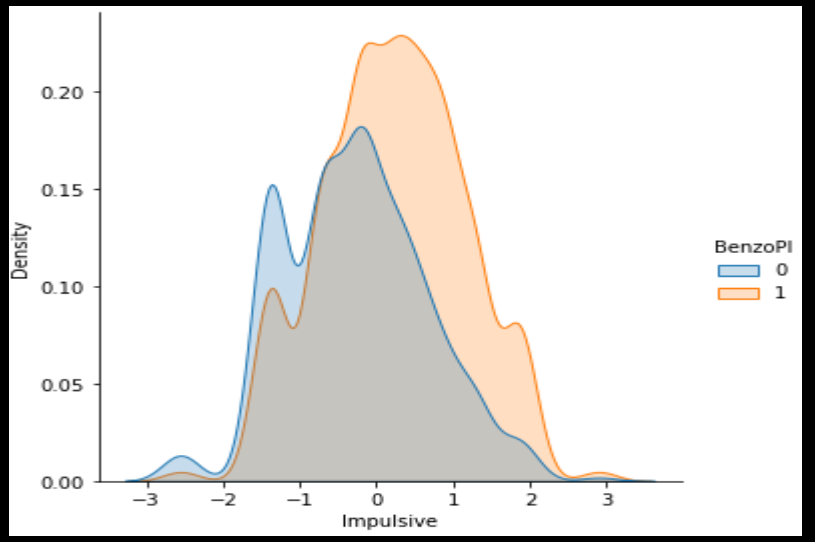




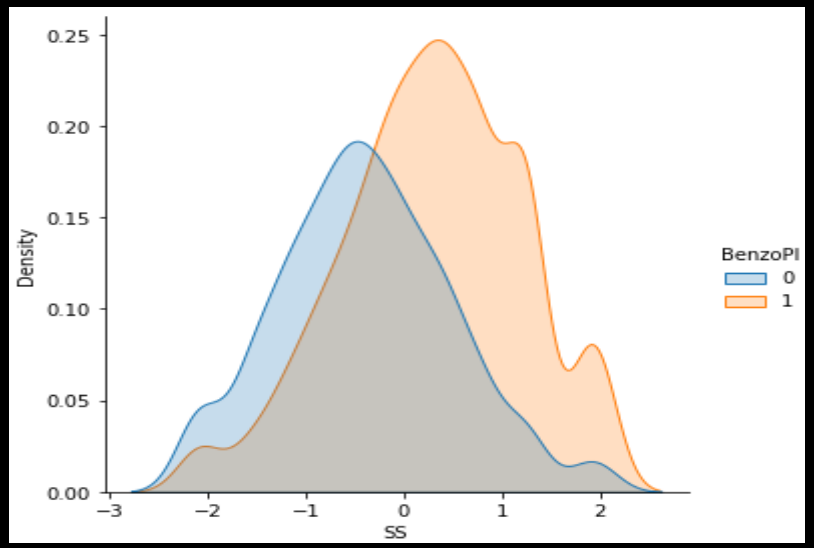


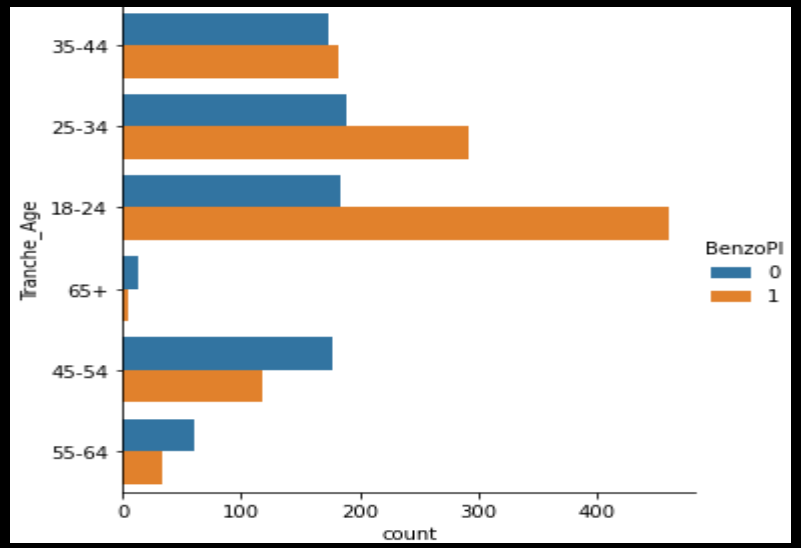


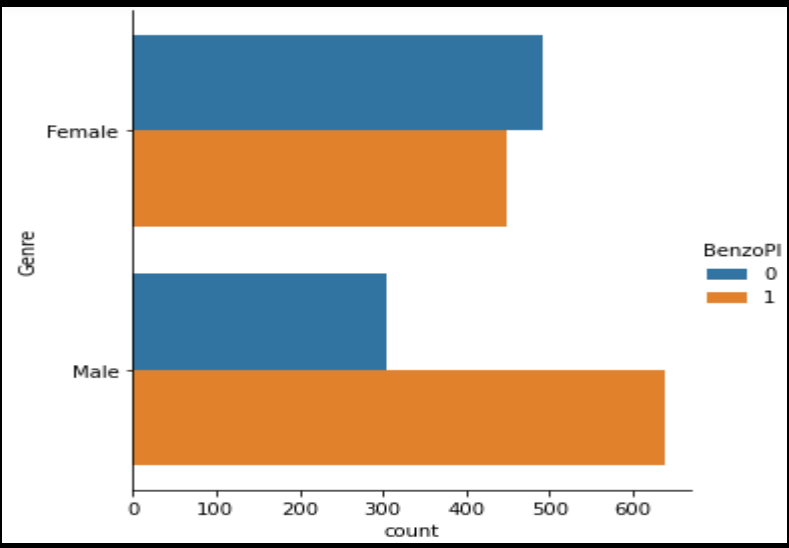


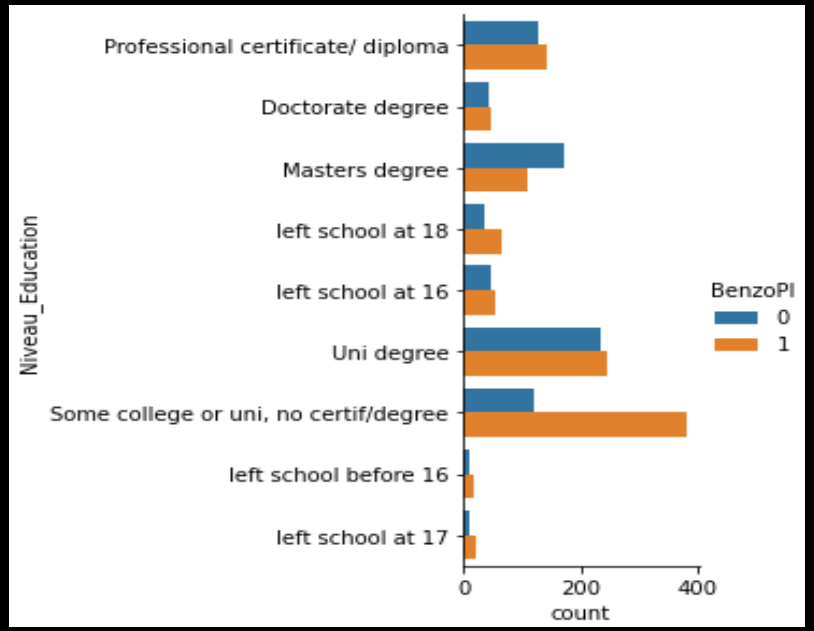


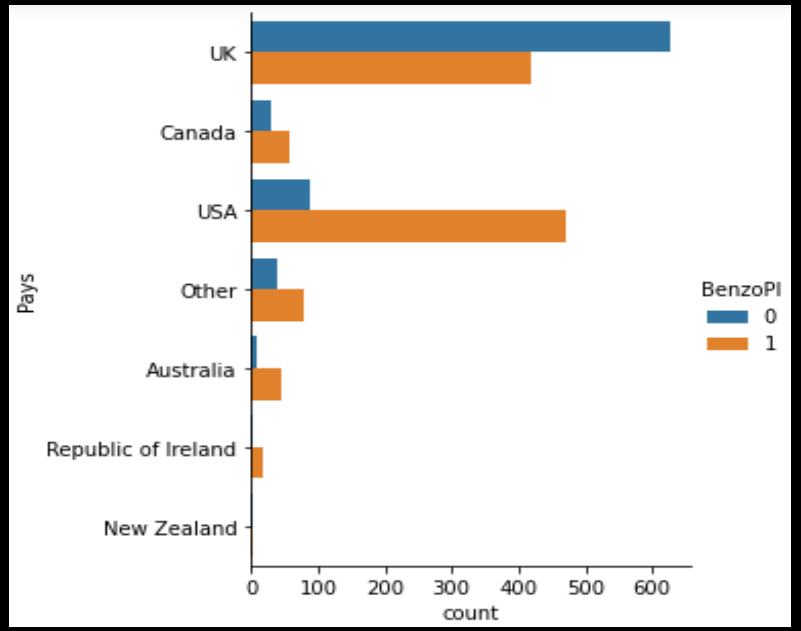
Visualisation des dépendances des features/targets (BenzoPl)

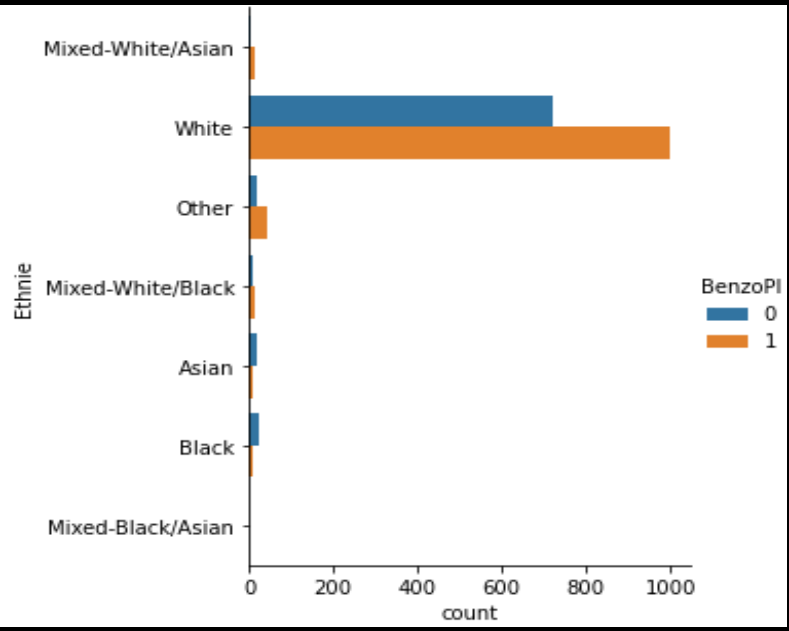








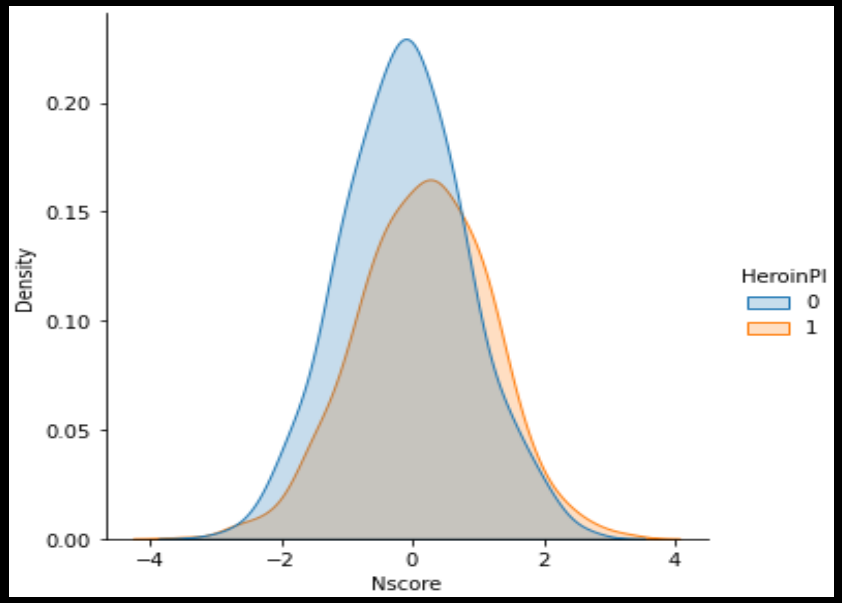


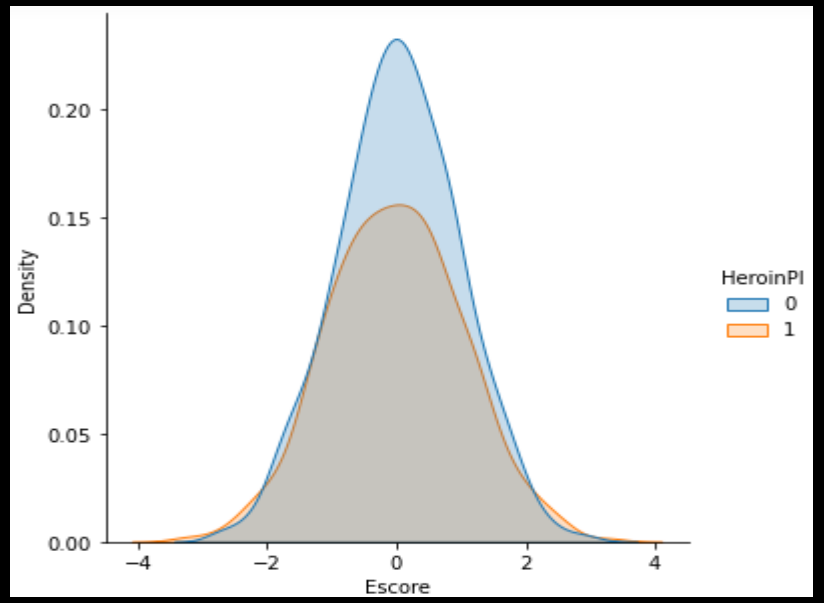


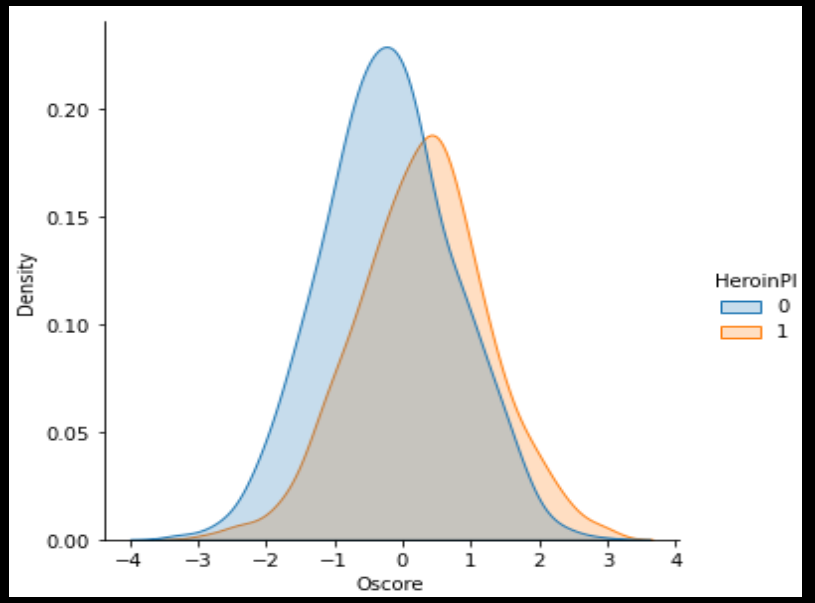
BenzoPl

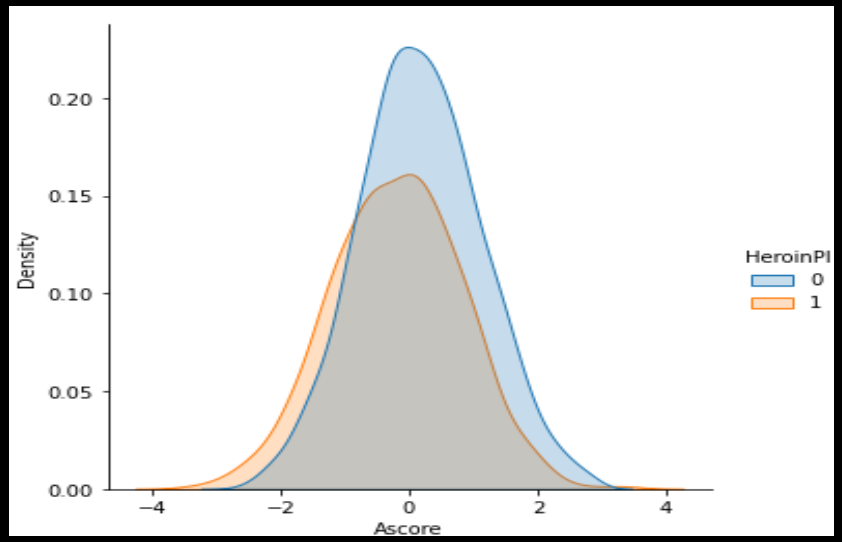
- En regardant le graphique de SS et celui de Impulsive, on peut facilement conclure que ces deux features sont très pertinentes pour l'étude de la consommation de BenzoPl en voyant les décalages.
- On peut également remarquer que les features pays et ethnie sont visiblement déséquilibrées donc pas très pertinentes.
- Les hommes ont plus tendance à consommer la pléiade BenzoPl que les femmes, de plus, il y a plus de femmes qui en consomment que celles qui n'en consomment pas.
- Les gens âgés entre 18 et 24 ans ont plus tendance à la consommer plus que les autres tranches d'âge.

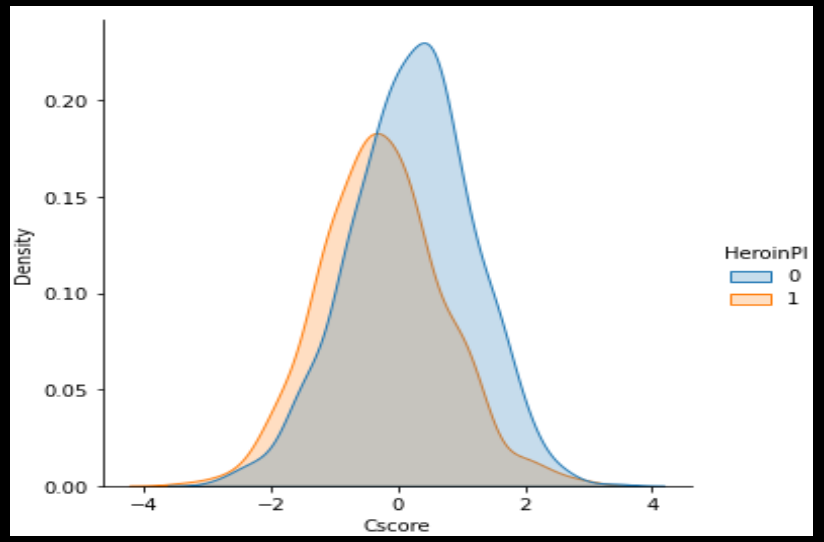
Visualisation des dépendances des features/targets (HeroinPl)

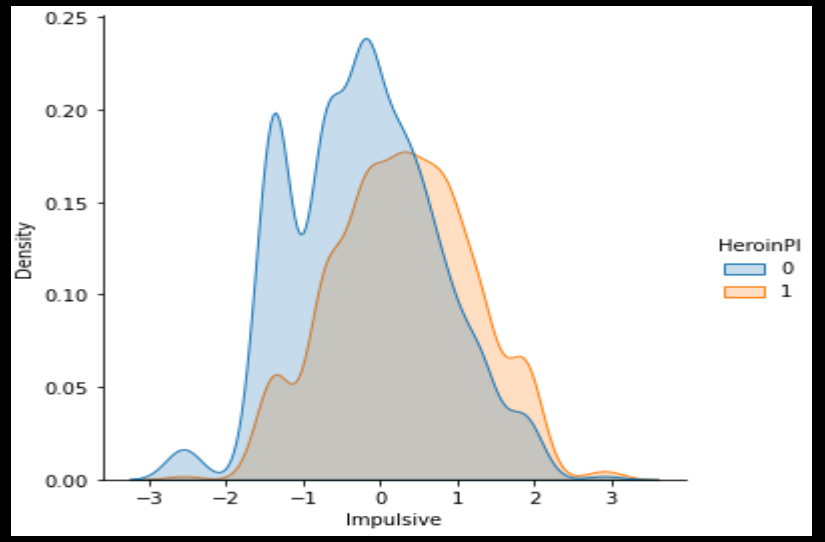




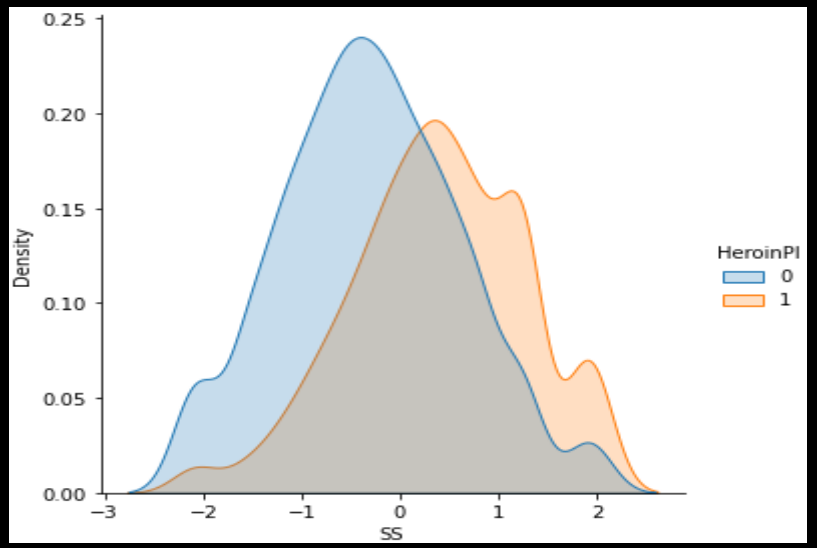


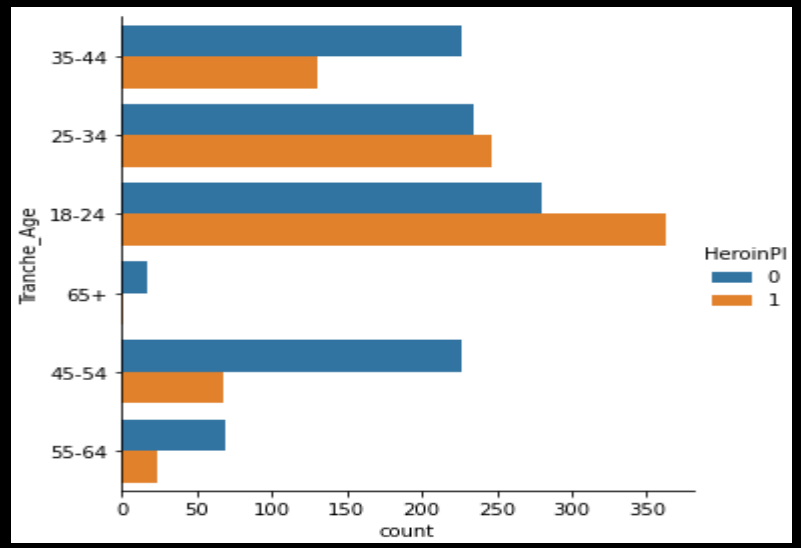


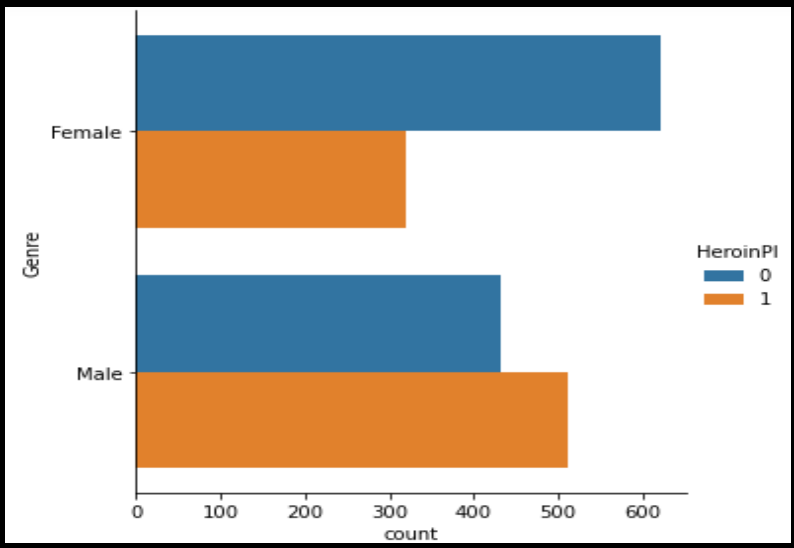


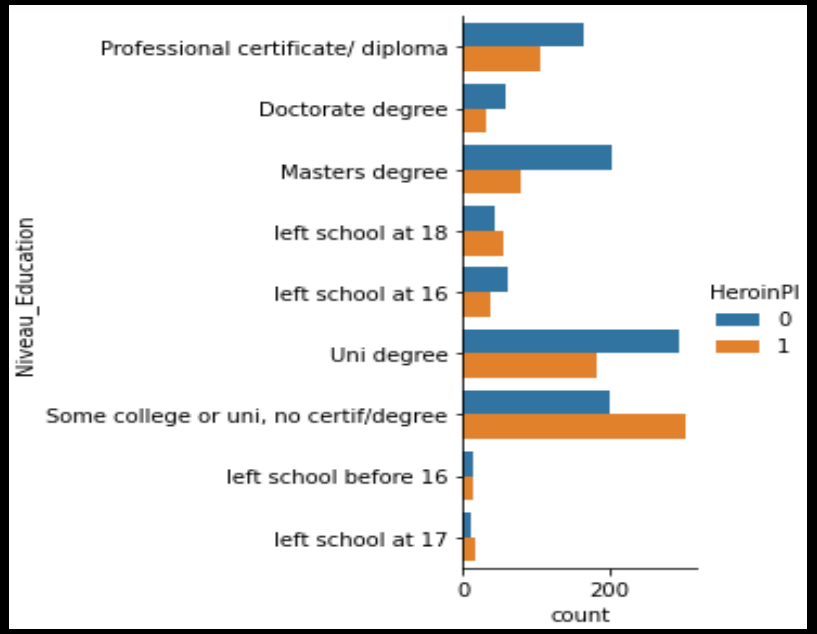


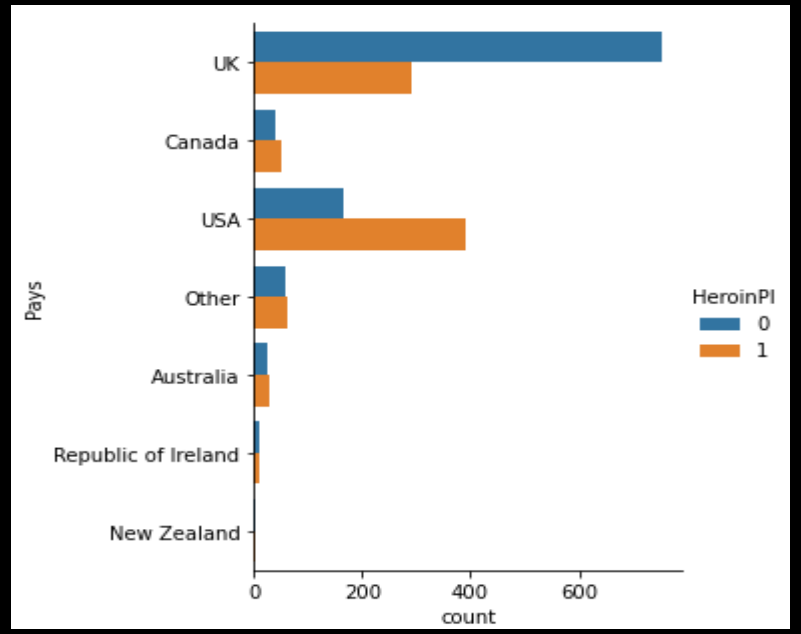
Visualisation des dépendances des features/targets (HeroinPl)

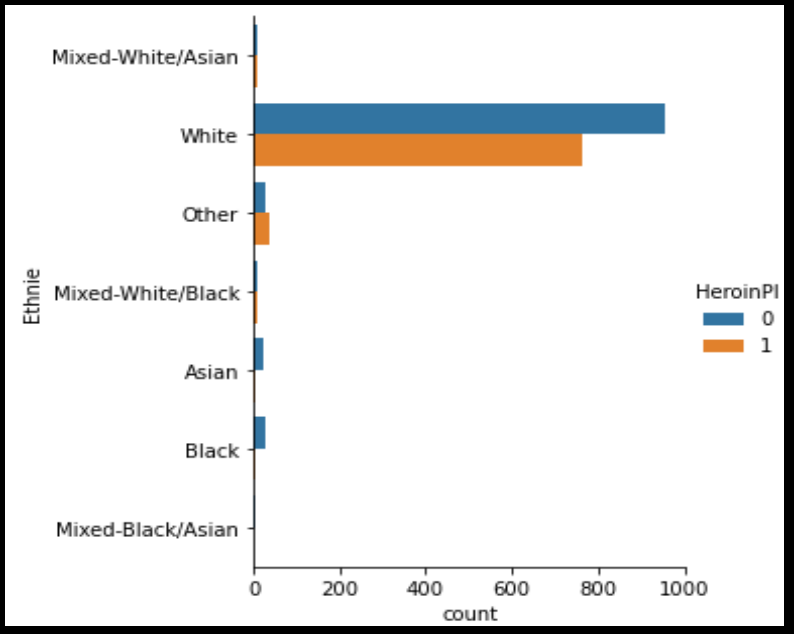








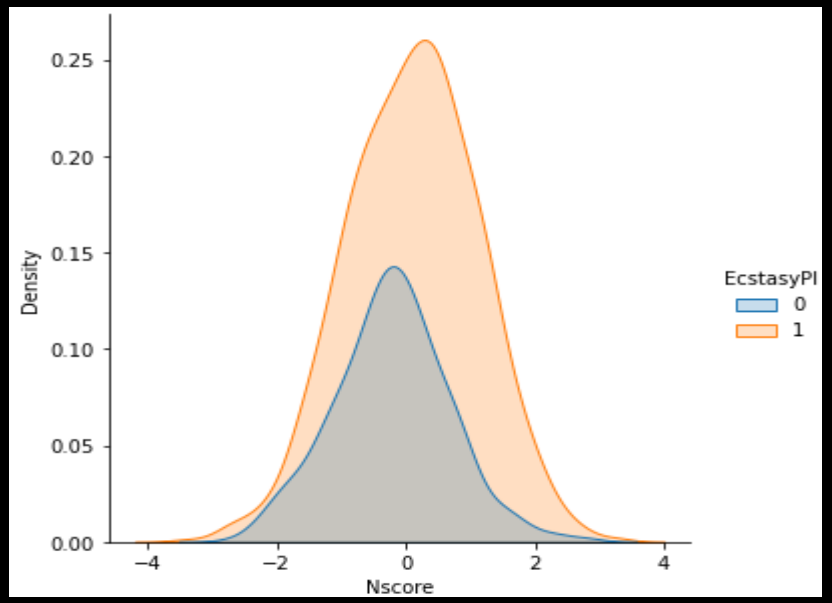


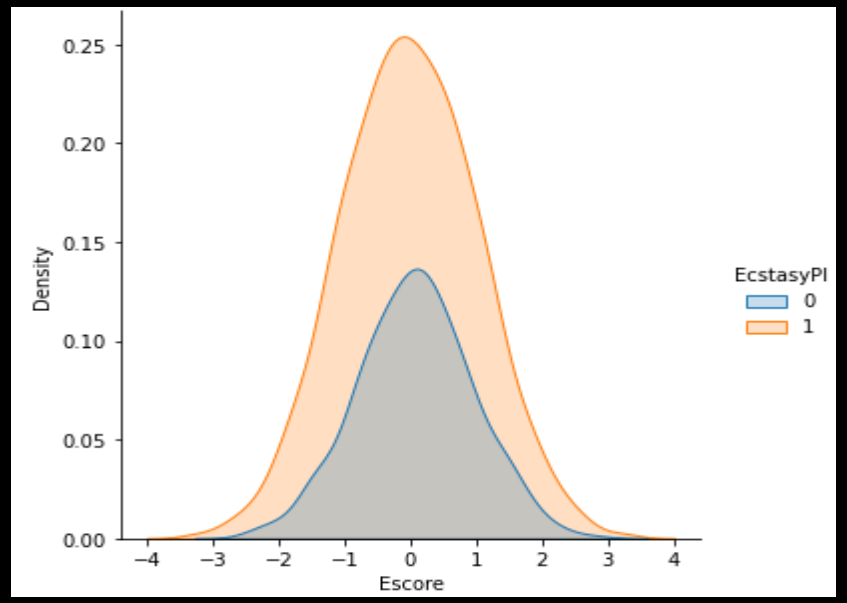


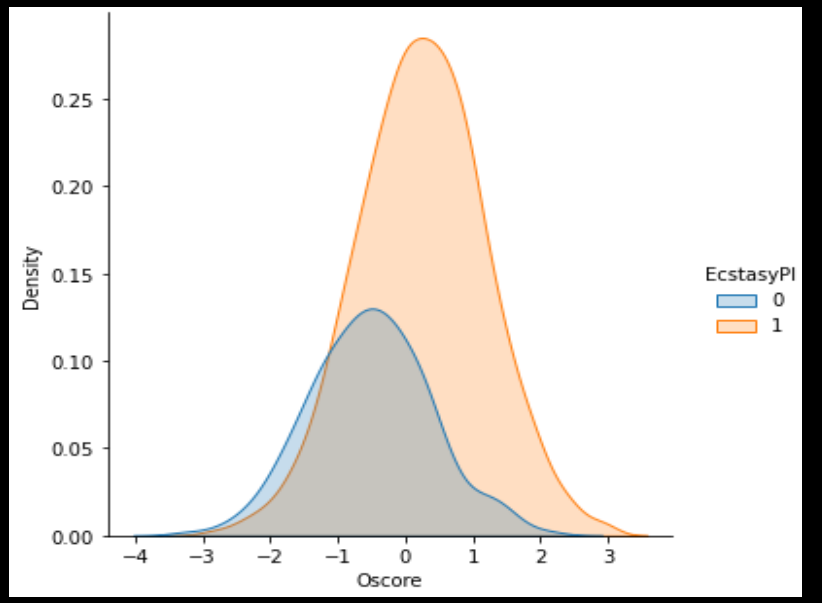
HeroinPl

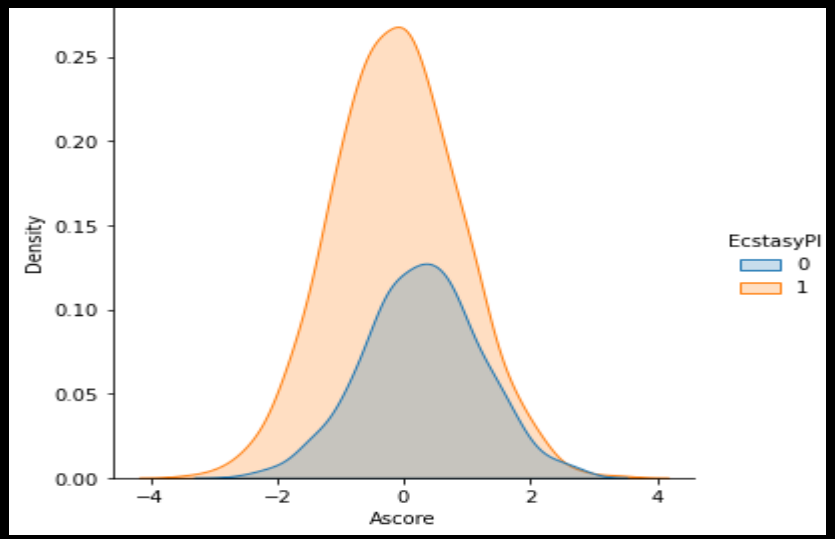
- Ici encore on peut voir que SS et Impulsive sont très pertinentes de par les décalages de leurs courbes.
- On remarque encore une fois que les features pays et ethnie sont visiblement déséquilibrées donc pas très pertinentes.
- Il y a beaucoup plus de femmes qui en consomment que celles qui n'en consomment pas.
- Les tranches d'âge 18-24 et 25-34 ont plus tendance à la consommer plus que les autres tranches d'âge.

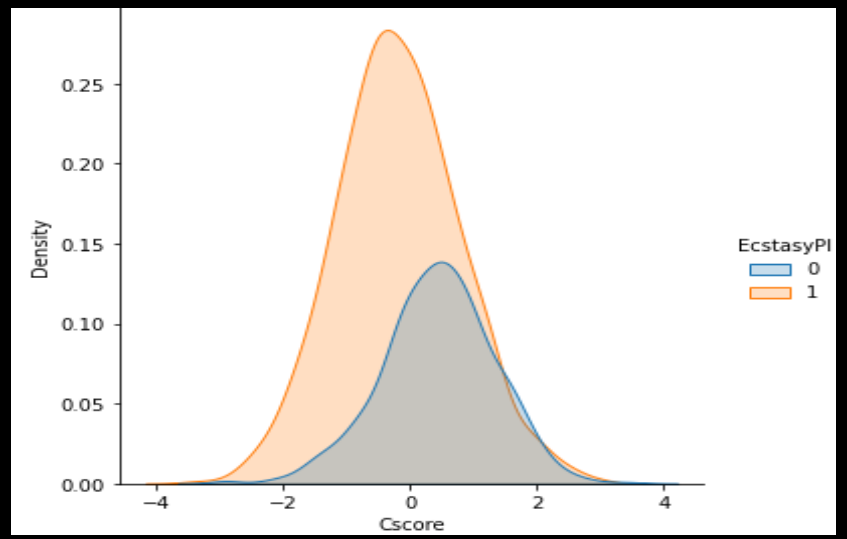
Visualisation des dépendances des features/targets (EcstasyPl)

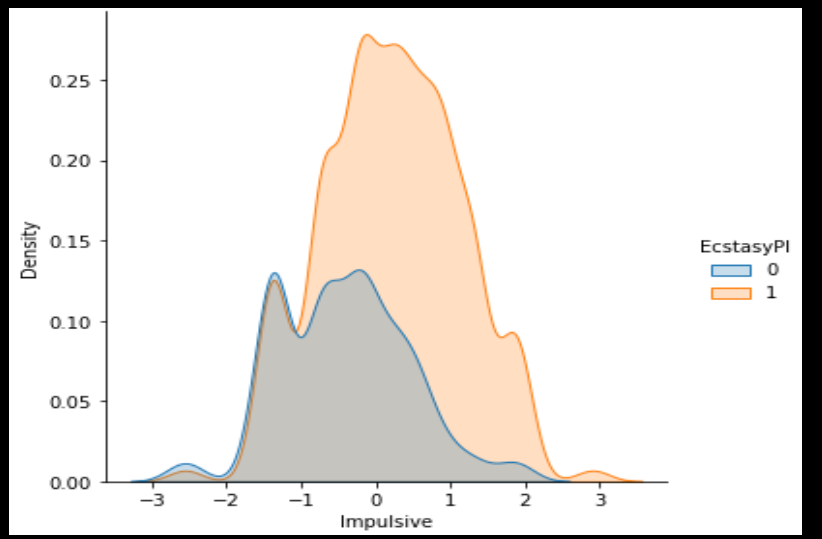




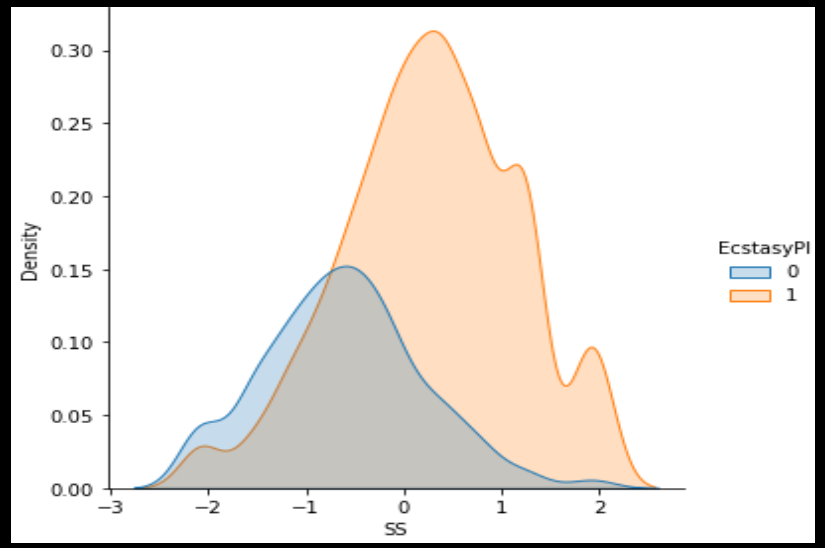


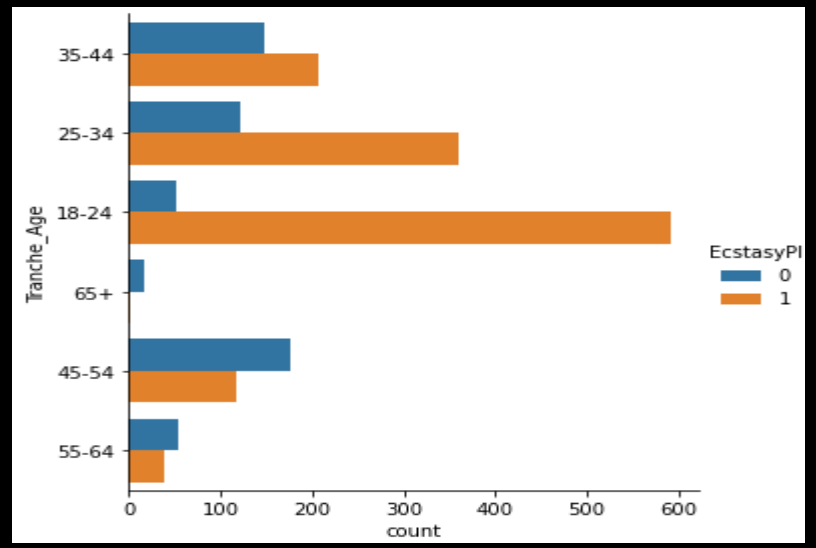


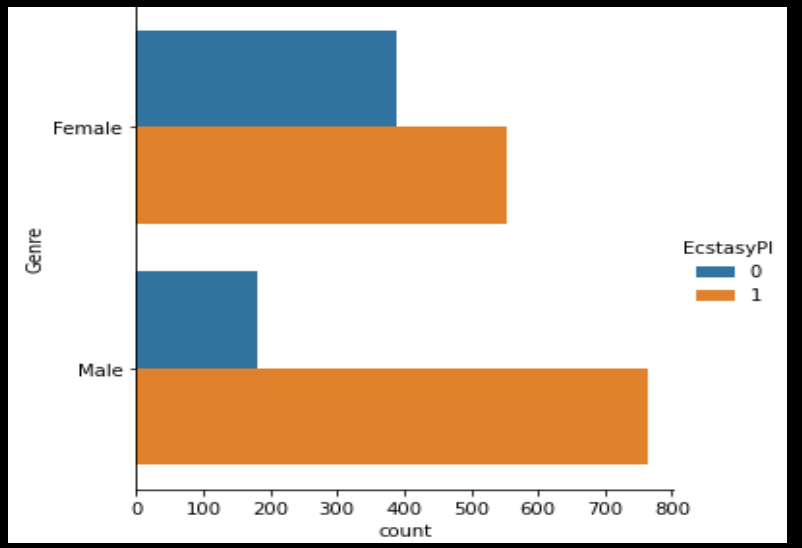


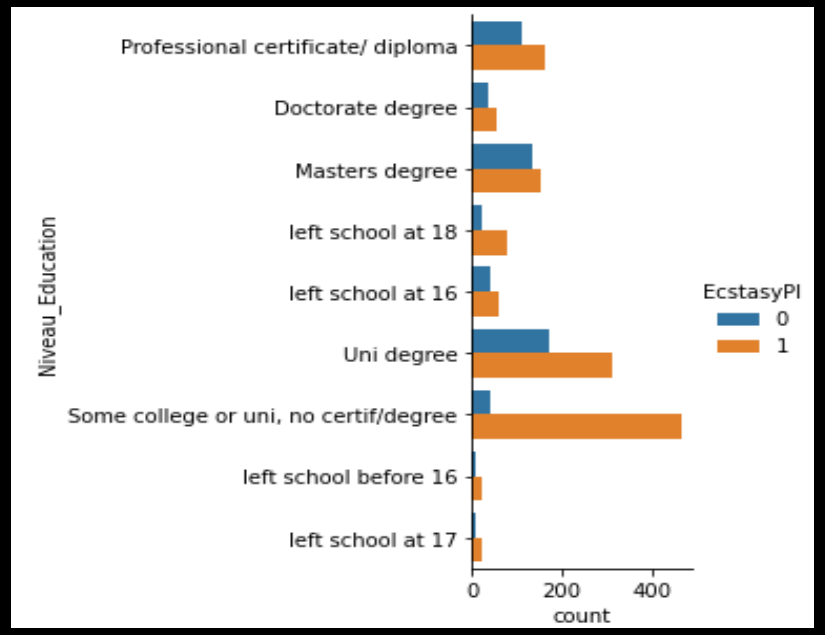


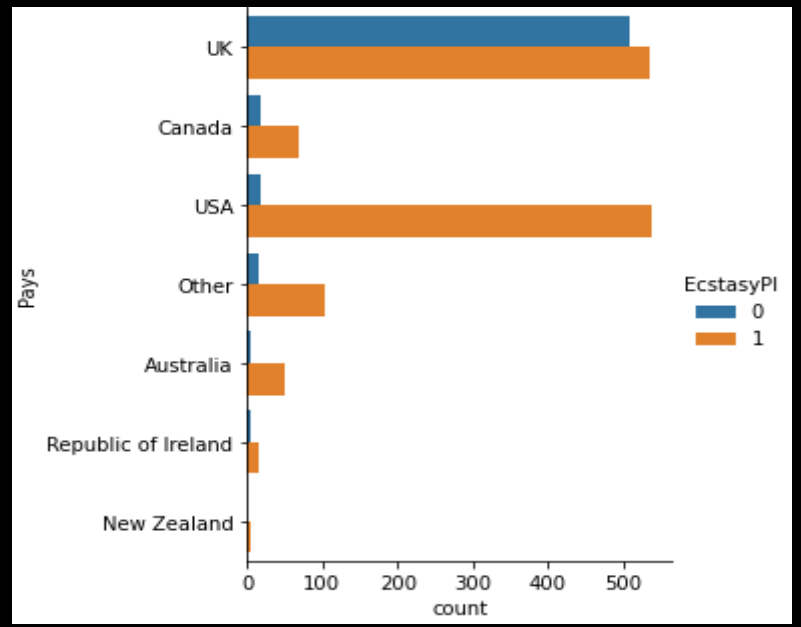
Visualisation des dépendances des features/targets (EcstasyPl)

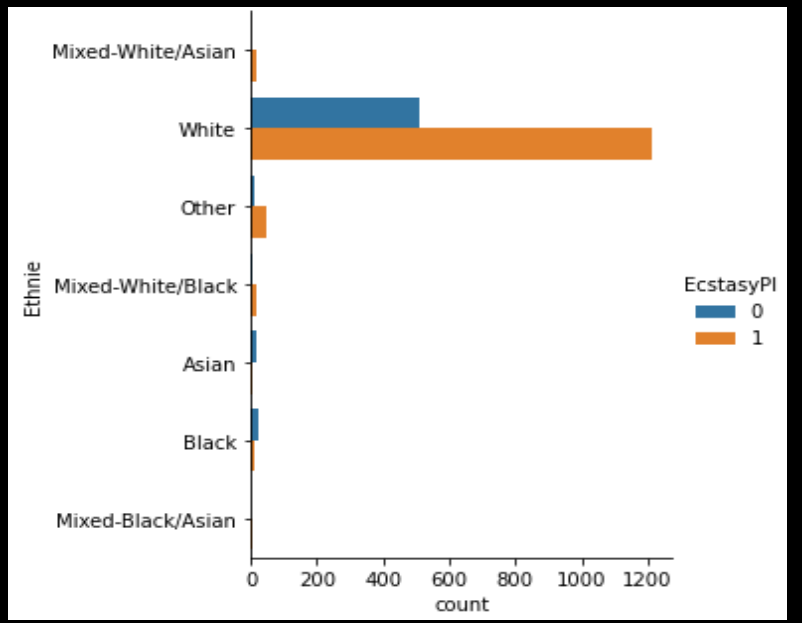












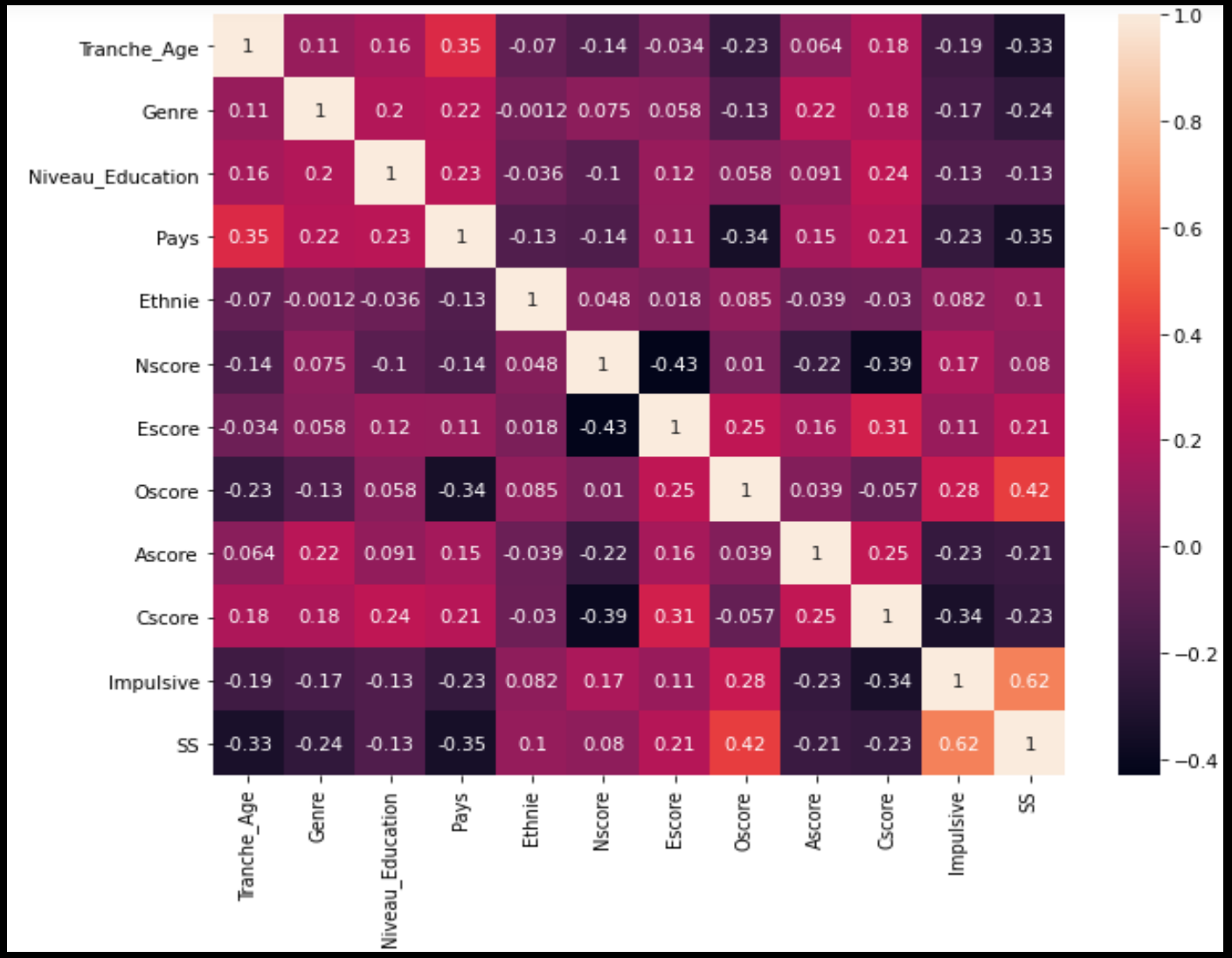
EcstasyPl

- Ici les courbes de SS et Impulsive ne sont pas aussi convaincantes que les deux autres pléiades, par contre la courbe de pays et celles d'ethnie confirment encore une fois à quel point ces deux features sont déséquilibrées et peu pertinentes.
- On remarque qu'il y a presque autant de consommateurs américains que des consommateurs anglais.
- Il n'y a quasiment aucun consommateur d'ecstasy parmi les personnes âgées de 65 ans ou plus.

Corrélation entre les features

- Dans les diapositives précédentes nous avons constaté que les features pays et ethnie n’étaient pas très pertinentes par rapport à la prédiction des targets. Mais qu’en est-il en vrai ?
- Jetons un coup d'œil sur la corrélation entre les différentes features pour vérifier comment elles interagissent entre elles.

Corrélations entre les features



- C’est très facile à conclure que la feature Ethnie n’est vraiment pas pertinente. Pays, ça serait à voir, mais pour que le critère soit significatif, il aurait fallu plus de données.

Pre-processing and Modelling

- On a vu précédemment que les features pays et ethnie ne sont pas pertinentes pour notre travail, alors on s'est permis de les retirer car elles n'apportent aucune information intéressante.
- La drogue fictive semeron, étant une drogue inventée pour repérer les sur-demandeurs, va aussi être retirée.
- On va considerer les colonnes Age, Niveau d'éducation et Genre, comme des catégories. On va les dummieser.

Pre-processing

Objectif: transformer le data pour le mettre dans un format propice au machine learning

- Création du Train Set / Test Set
- Encodage

Modelling

Objectif: développer un modèle de machine learning capable de répondre a l'objectif final.

- Définir des fonctions utiles au modelling
- Selection des meilleurs modèles
- Optimisation avec GridSearchCV
- Analyse des modèles, learning curve et prise de décision

Modelling

- Nous avons créé une fonction pour nous retourner l'évaluation des modèles que nous voulons tester. Ici nous utilisons le f1 score étant donné que l'EcstasyPl est peu équilibré, c'est beaucoup plus pertinent que l'accuracy dans ce cas.
- Ensuite nous avons défini une fonction qui analyse les meilleurs modèles pour chaque drogue du dataset de base. Parmi ces modèles nous avons le Random Forest, le KNN, les arbres de décision, le Naive Bayes, l'analyse discriminante linéaire, la regression logistique et le SVM.
- Après avoir exécuté cette fonction sur les drogues du dataset on a pu l'utiliser également pour les différentes pléiades.

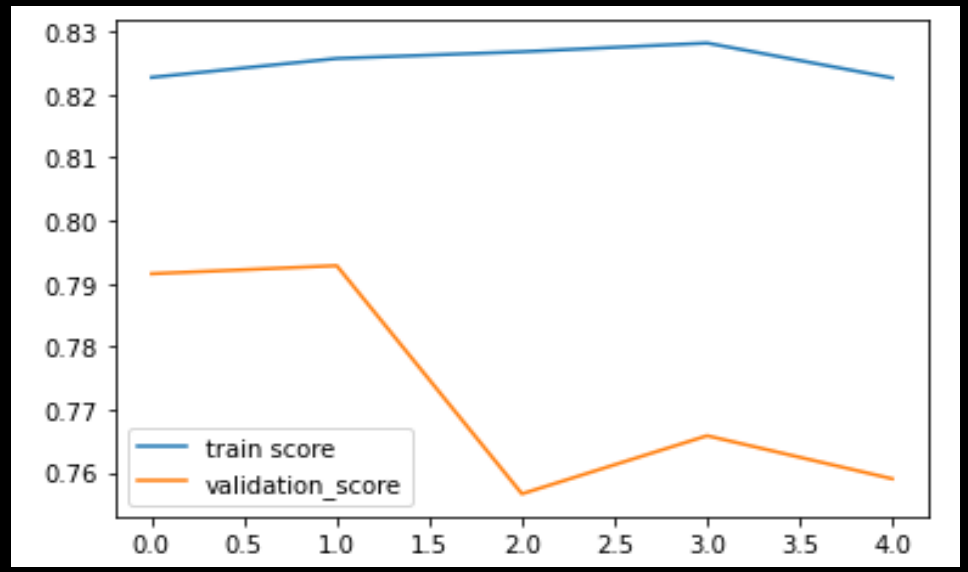
Meilleurs modèles pour chaque pléiade

1. Pour l'HeroinPl le random forest donne un meilleur modèle.
2. Pour la BenzoPl et l'EcstasyPl c'est le SVM qui est le meilleur modèle

Maintenant il nous reste plus qu'à améliorer chaque modèle trouvé afin d'avoir des résultats optimaux.

BenzoPl

1. Courbe d'entraînement et de validation (cross validation)



Nous avons un F1 score de 0,83 pour l'entraînement et 0,77 pour la validation en moyenne sur la cross validation.

2. Matrice de confusion du train et test set

i. Train set

	0	1
0	409	219
1	113	767

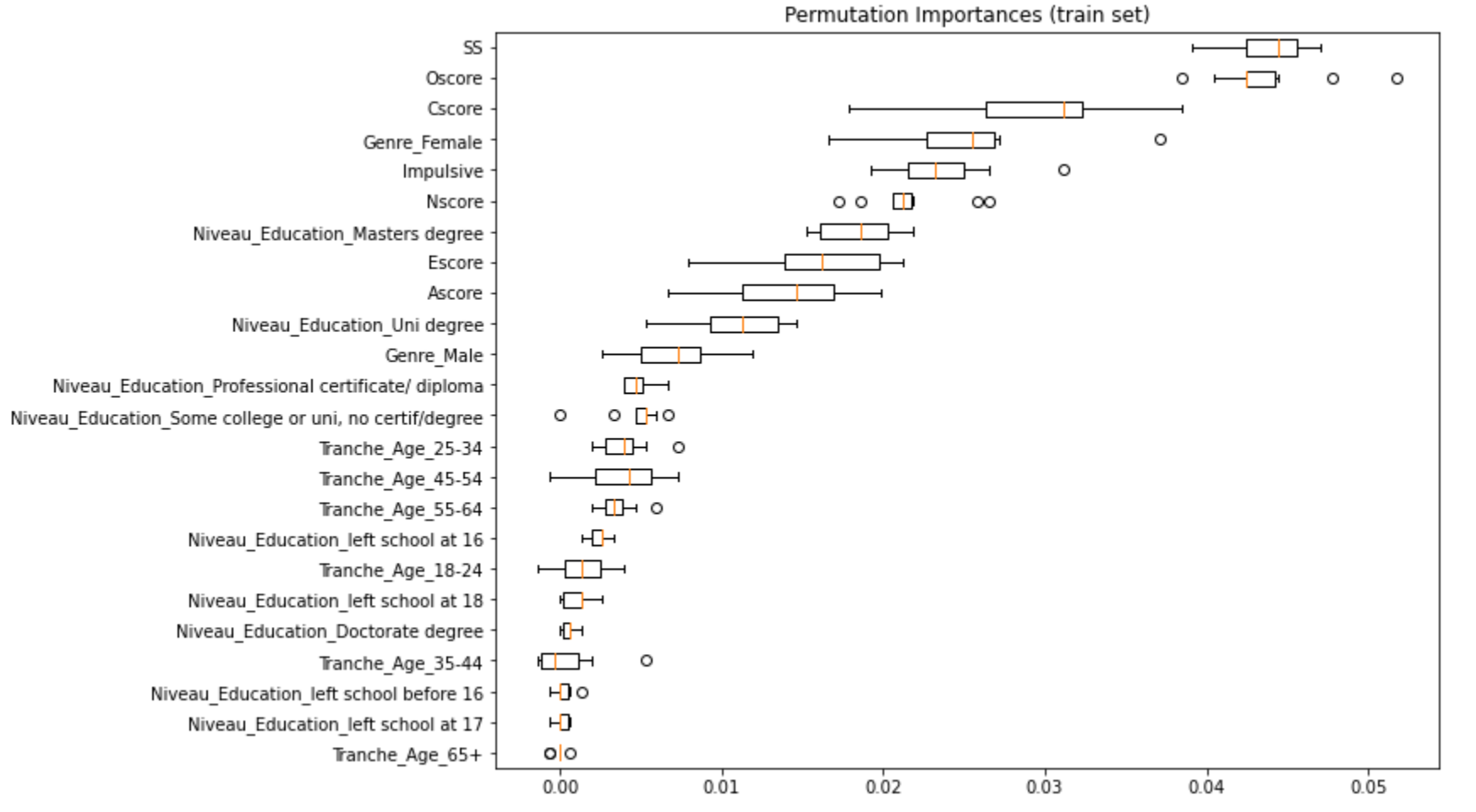
F1 score : 0.82
Recall score 0.87

ii. Test set

	0	1
0	107	61
1	43	166

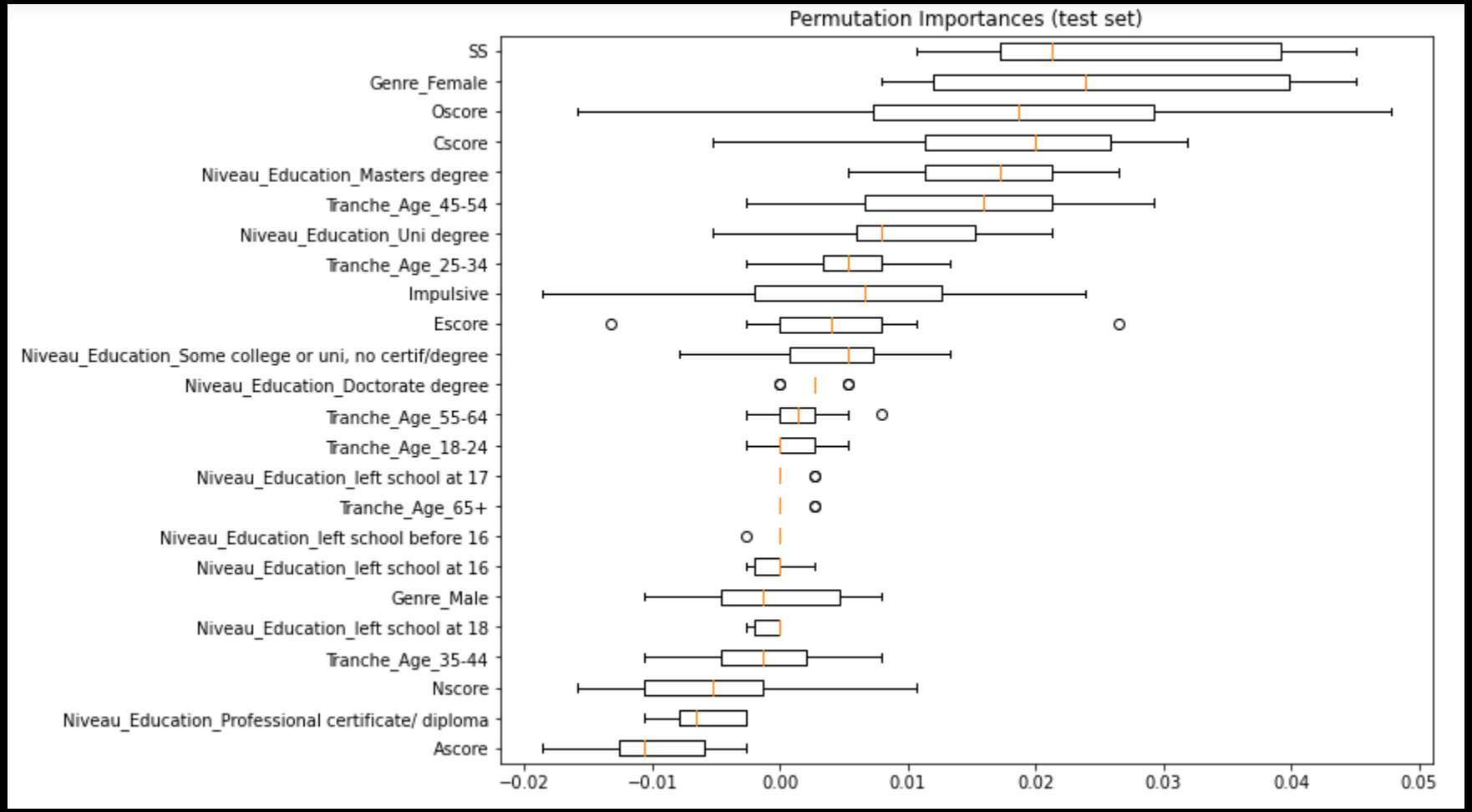
F1 score : 0.76
Recall score 0.79

Maintenant nous allons voir l'ordre d'importance de chaque feature pour la BenzoPl avec SVM(train set)



On voit que la recherche de sensation (SS) et Oscore sont des features très importantes pour l'étude de la consommation de la BenzoPl

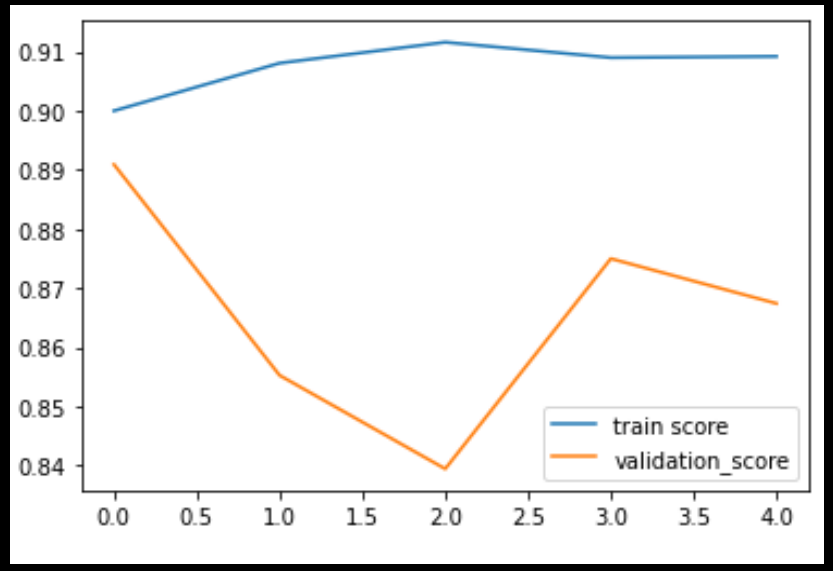
L’ordre d’importance de chaque feature pour la BenzoPl avec SVM(test set)



L’ordre d’importances est quasiment le même que pour le train set, mais nous avons de plus grands box pour le test. Donc il y a une de plus grandes variances.

EcstasyPl

1. Courbe d'entraînement et de validation (cross validation)



Nous avons un F1 score de 0,91 pour l'entraînement et 0,87 pour la validation en moyenne sur la cross validation.

2. Matrice de confusion du train et test set

i. Train set

	0	1
0	320	127
1	67	994

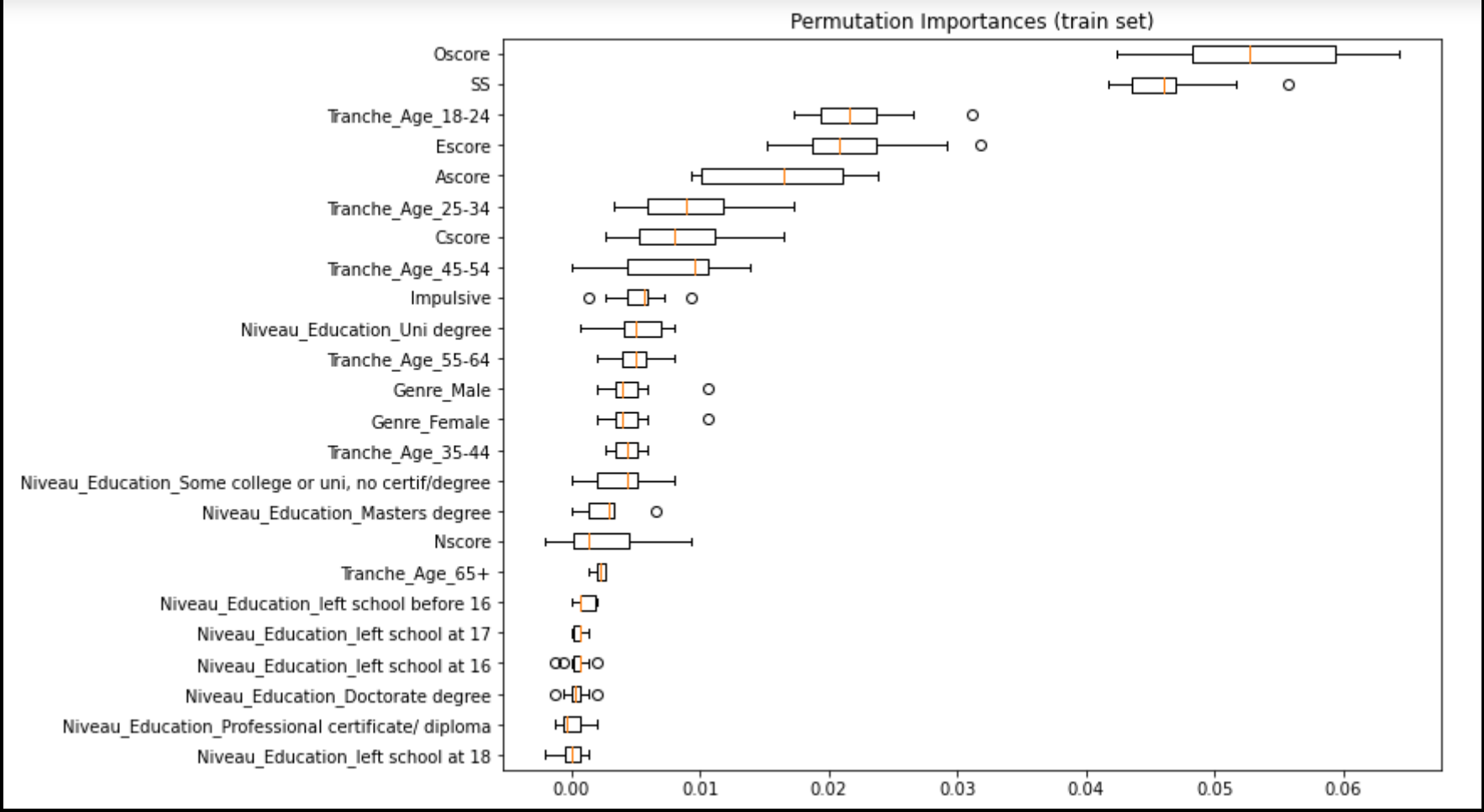
F1 score : 0.91
Recall score 0.94

ii. Test set

	0	1
0	86	35
1	28	228

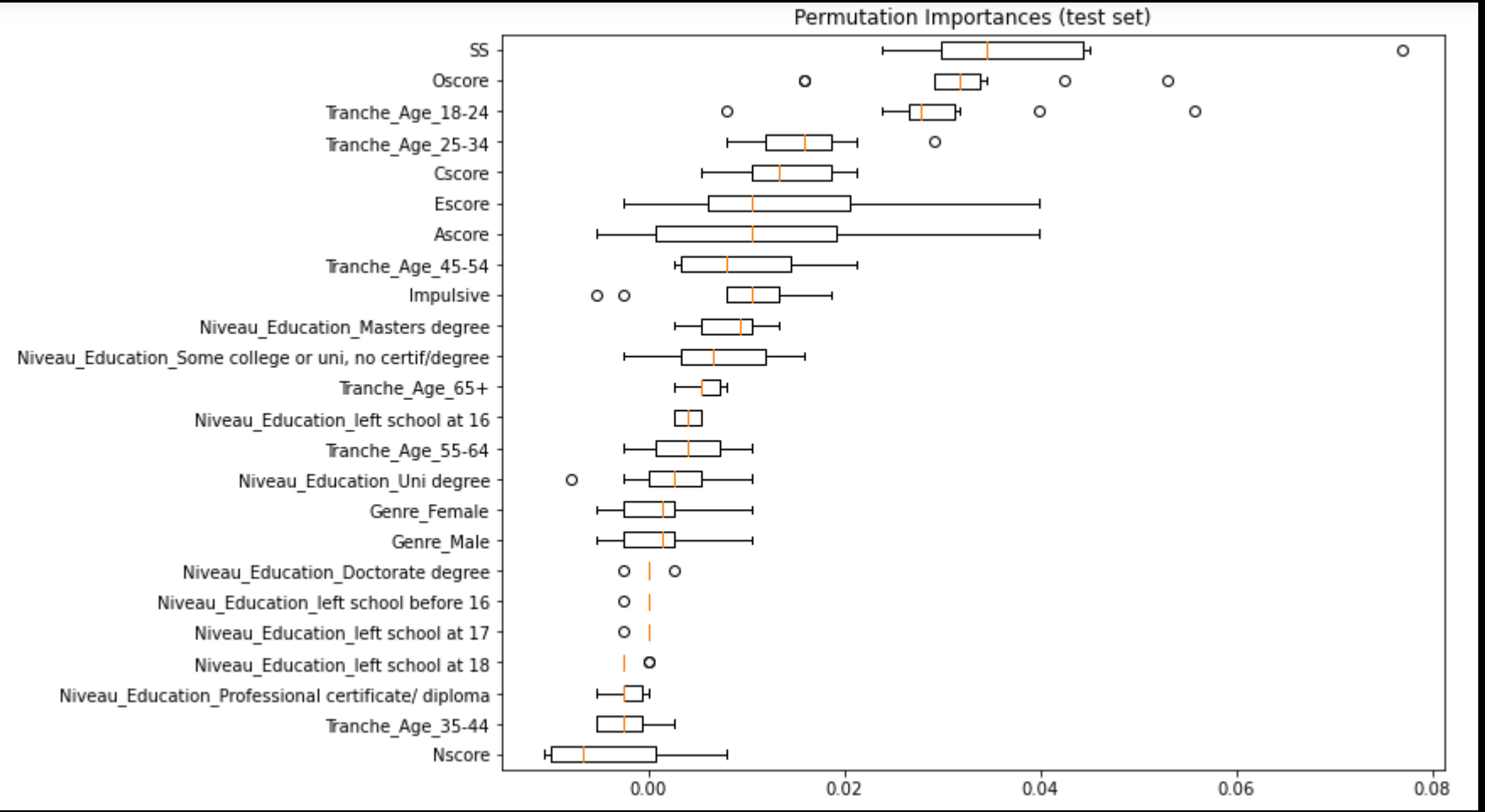
F1 score : 0.88
Recall score 0.89

Maintenant nous allons voir l'ordre d'importance de chaque feature pour l'EcstasyPl avec SVM(train set)



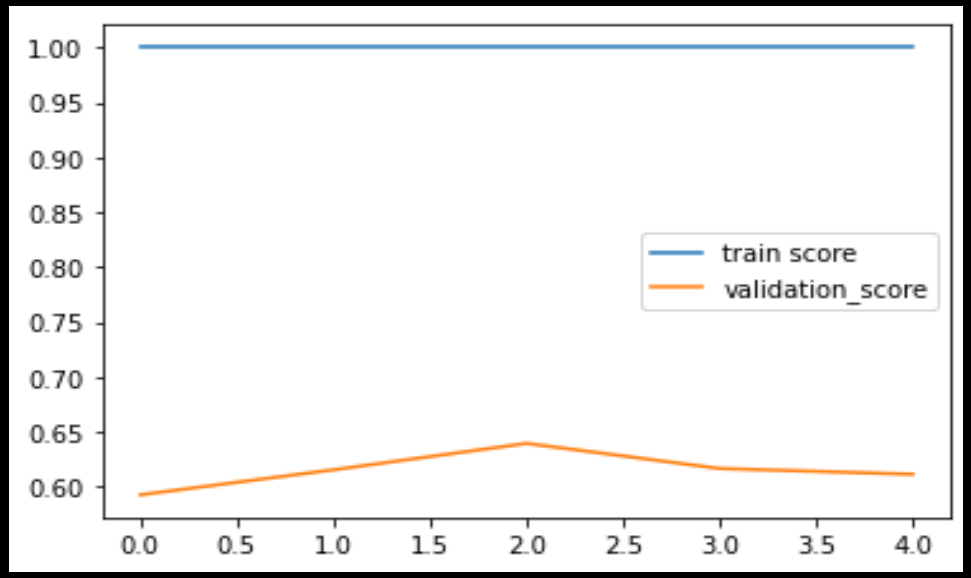
Une fois de plus la recherche de sensation (SS) et Oscore sont des features très importantes, cette fois pour l'étude de la consommation de l'EcstasyPl.

L’ordre d’importance de chaque feature pour l’EcstasyPl avec SVM(test set)



HeroinPl

1. Courbe d'entraînement et de validation (cross validation)



Nous avons un F1 score de 1 pour l'entraînement et 0,62 pour la validation, pas très fameux.

2. Matrice de confusion du train et test set

i. Train set

	0	1
0	824	0
1	0	684

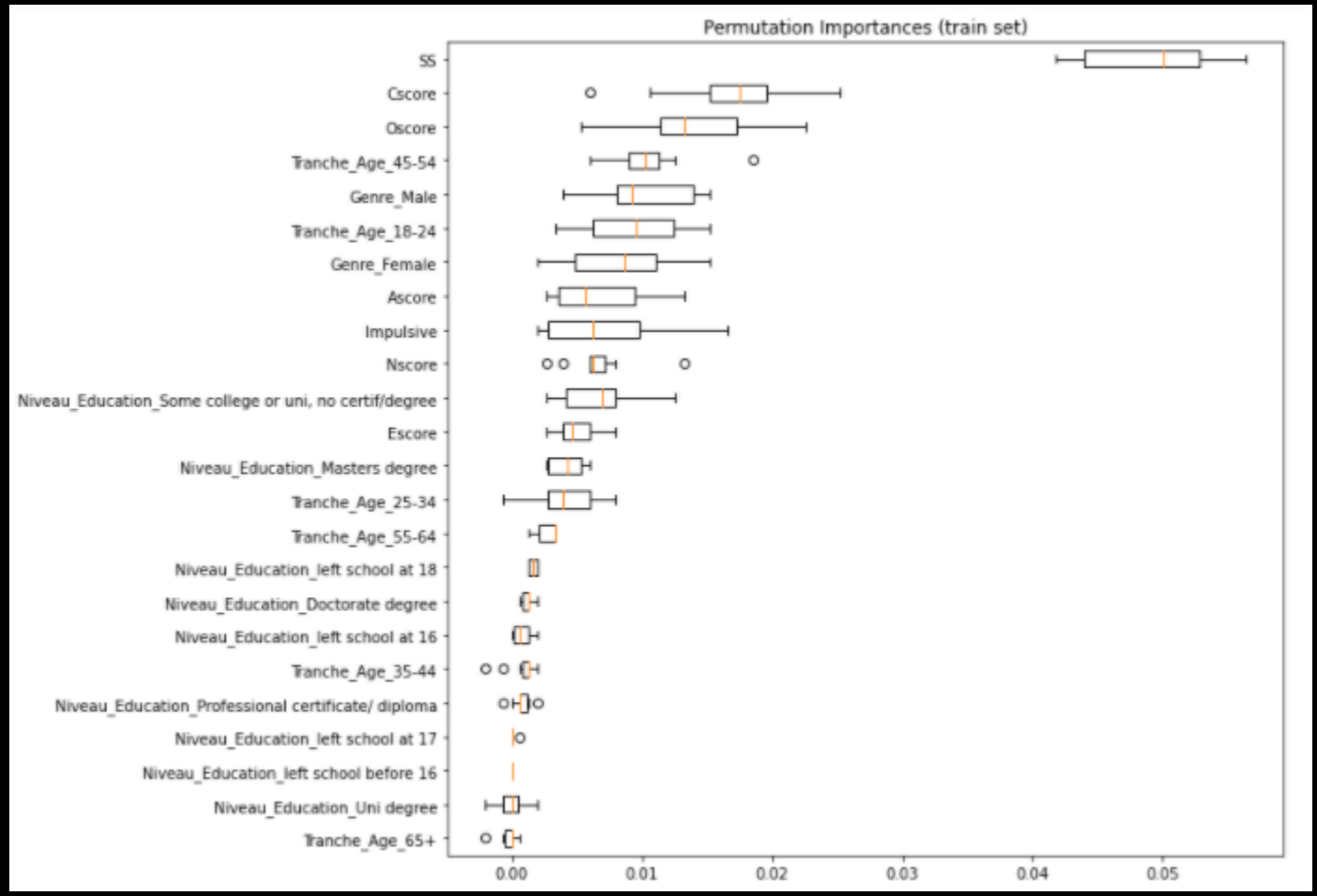
F1 score : 1.00
Recall score 1.00

ii. Test set

	0	1
0	176	53
1	40	108

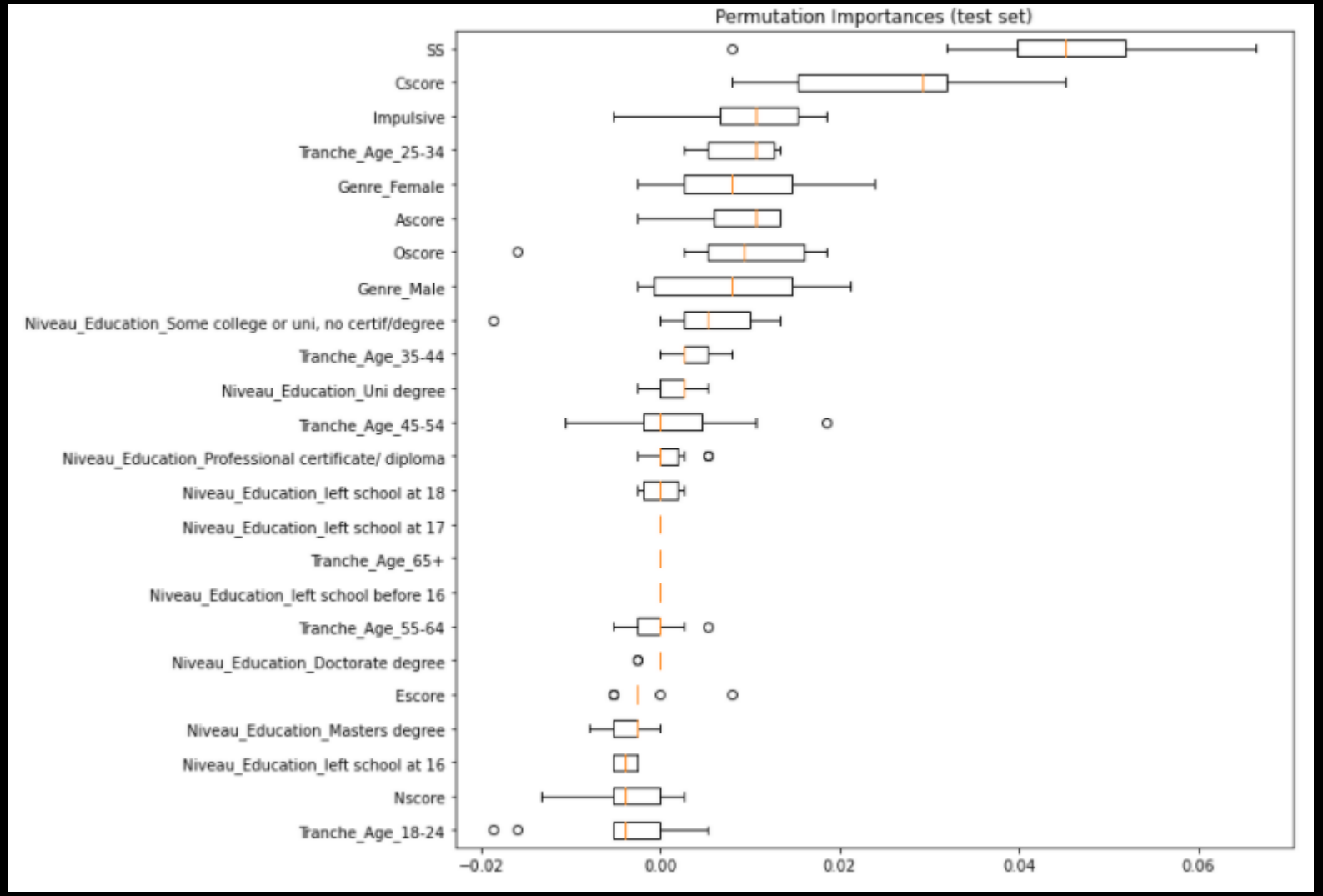
F1 score : 0.70
Recall score 0.73

Maintenant nous allons voir l'ordre d'importance de chaque feature pour l'HeroinPl avec SVM(train set)



On voit que la recherche de sensation (SS), Cscore et Oscore sont des features très importantes pour l'étude de la consommation de la BenzoPl

L'ordre d'importance de chaque feature pour l'HeroinPl avec SVM(test set)



- On voit que la recherche de sensation (SS) et Cscore sont des features très importantes pour l'étude de la consommation de l'HeroinPl.

Pourquoi avoir choisi le recall ?

- Recall = TruePositives / (TruePositives + FalseNegatives)
- Nous pouvons voir dans la formule que moins il y a de faux négatifs, plus le recall est grand et vice-versa.
- En essayant d'augmenter la valeur du recall, nous baissons les faux négatifs et cela est bon pour l'algorithme sur lequel nous travaillons étant donné que c'est un algorithme préventif. Il est préférable de prédire plus de faux positifs que de faux négatifs.
  - En effet, si l'on se trompe en prédisant qu'une personne donnée est susceptible de consommer de la drogue, le pire qui puisse arriver serait peut-être qu'elle puisse passer des heures dans une salle pour désintoxication.
  - Par contre si on se trompe en prédisant qu'une personne ne consomme pas la drogue alors qu'en vraie cette personne en consomme elle risque de détruire sa vie sans que nous ne puissions intervenir.

Utilisation de l’API

- Nous avons pu récupérer nos modèles et nous les avons enregistrés dans des fichier pkl pour pouvoir les utiliser dans le code Flask qui nous a permis de créer l’API.

- Grâce à HTML, CSS et Java Script nous avons pu créer les pages internet sur lesquelles tester le fonctionnement de nos modèles.

home.html

Drugs consumption prediction

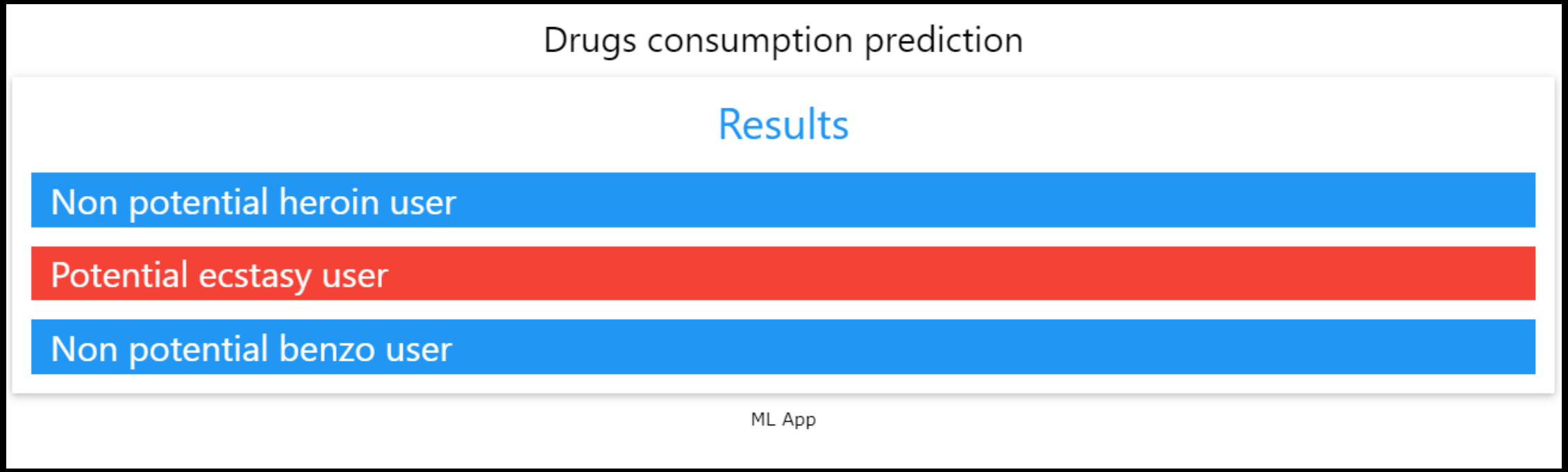
Please enter the data in the following boxes.

Tranche age; value min : 0 and value max 5	0
Genre	Male
Niveau Education; value min : 0 and value max 8	0
Nscore; value min : 0 and value max 48	0
Escore; value min : 0 and value max 41	0
Oscore; value min : 0 and value max 34	0
Ascore; value min : 0 and value max 40	0
Cscore; value min : 0 and value max 40	0
Impulsive; value min : 0 and value max 9	0
SS; value min : 0 and value max 10	0

predict

Machine Learning App with Flask

Result.html


Drugs consumption prediction
Results
Non potential heroin user
Potential ecstasy user
Non potential benzo user
ML App

Conclusion

- En conclusion de cette étude, nous pensons que les critères utilisés dans la prédiction de la consommation de drogues sont relativement pertinents. Néanmoins, pour obtenir de bonne performance dans la prédiction de la consommation de drogues, d'autres paramètres doivent surement rentrer en compte comme la localisation, par exemple. Dans la majorité des modèles implémentés, plus de données permettrait d'augmenter les performances.

- Nous vous déconseillons vivement d’essayer toutes ces combinaisons de drogues que nous venons d'étudier.