Etude d’un dataset sur la consommation de drogue: classification des personnes ayant des disposition à devenir consommateur

Matthias PICARD
Severin LEFEBURE

Objectifs de notre étude

- Nous travaillons pour une association ou un organisme d'assistance sociale qui souhaite anticiper si une personne a plus ou moins de chance d’être ou de devenir consommateur régulier de drogue.

- Nous allons résoudre ce problème en développant un modèle de prédiction qui estimera une probabilité qu’une personne soit dans un cas ou dans l’autre.

- Nous considérons qu'il est plus grave que le modèle détermine qu'une personne ne consomme pas de drogue alors que c'est le cas (faux négatif), que le cas inverse où l'algorithme est trop prudent et considère qu'une personne avec peu de risque en consomme (faux positif). Dans notre cadre, il faut détecter tous les cas à risque même si cela implique d’être moins précis.

- C’est pourquoi notre objectif sera de viser un « recall » de 80% c'est à dire que le modèle classe au moins 80% des vraies potentiels toxicomanes en tant que tel, et n'en oublie que 20%. Le but sera de limiter au maximum les conséquences sur la « précision » du modèle quand on fixe cette valeur (c'est à dire limiter le nombre de personnes considérées comme toxicomane alors qu'elles sont "safe").

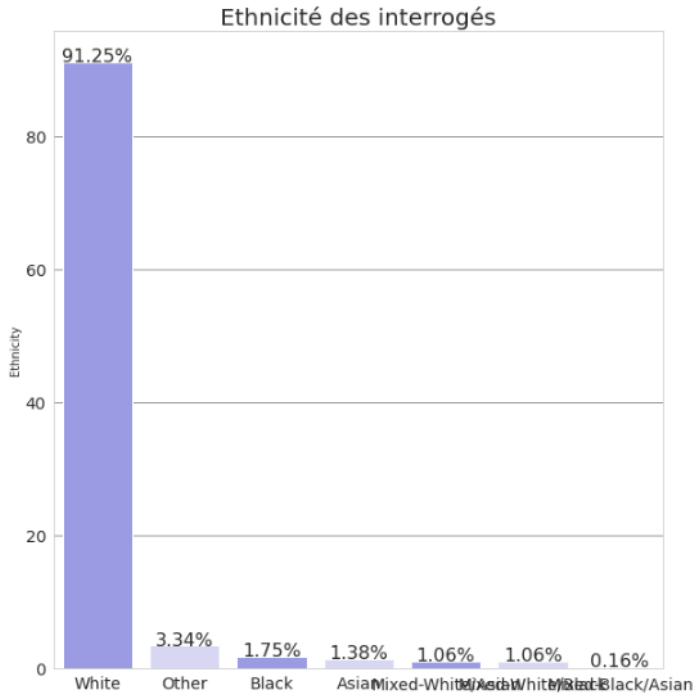
Présentation du dataset

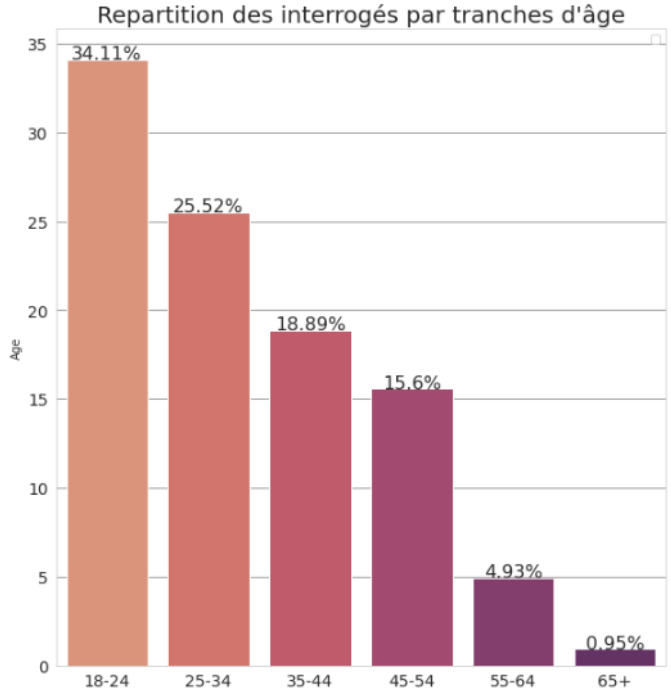
- Dataset issu d'un sondage anonyme sur internet, actif entre mai 2011 et mars 2012, crée par le Nottinghamshire Healthcare NHS Trust et qui fut utilisé pour un papier scientifique de 2015 qui étudiait la consommation de drogue en fonction notamment de la personnalité des individus.
- 1885 interrogés et 31 données recueillies pour chacun d'entre eux:
  - 5 informations personnelles: âge, genre, niveau d'éducation, pays d'origine et ethnicité
  - 7 traits de personnalités: « anxiété », « niveau d'extraversion », « ouverture à l'expérience », « compassion » , « niveau rigueur », « impulsivité » et « recherche de sensation », évalué avec un score.
  - Des informations sur la date de leur dernière consommation de drogues parmi 19 drogues, dont une fictive (le Semeron) pour repérer les participants susceptibles de mentir dans leurs réponses.
- Nous avons retiré les personnes ayant indiquées qu'elles avaient consommées du Semeron.
- Aucun NaN, ni de données dupliquées.

	Age	Gender	Education	Country	Ethnicity	Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	...	crack	ecstasy	heroin	ketamine	legal highs	LSD	methadone	mushrooms	nicotine	volatile substance abuse
0	35-44	Female	Professional certificate/ diploma	UK	Mixed-White/Asian	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	...	0	0	0	0	0	0	0	0	2	0
1	25-34	Male	Doctorate degree	UK	White	-0.67825	1.93886	1.43533	0.76096	-0.14277	...	0	4	0	2	0	2	3	0	4	0
2	35-44	Male	Professional certificate/ diploma	UK	White	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	...	0	0	0	0	0	0	0	1	0	0
3	18-24	Female	Masters degree	UK	White	-0.14882	-0.80615	-0.01928	0.59042	0.58489	...	0	0	0	2	0	0	0	0	2	0
4	35-44	Female	Doctorate degree	UK	White	0.73545	-1.63340	-0.45174	-0.30172	1.30612	...	0	1	0	0	1	0	0	2	2	0

Analyse du Dataset

Données personnelles



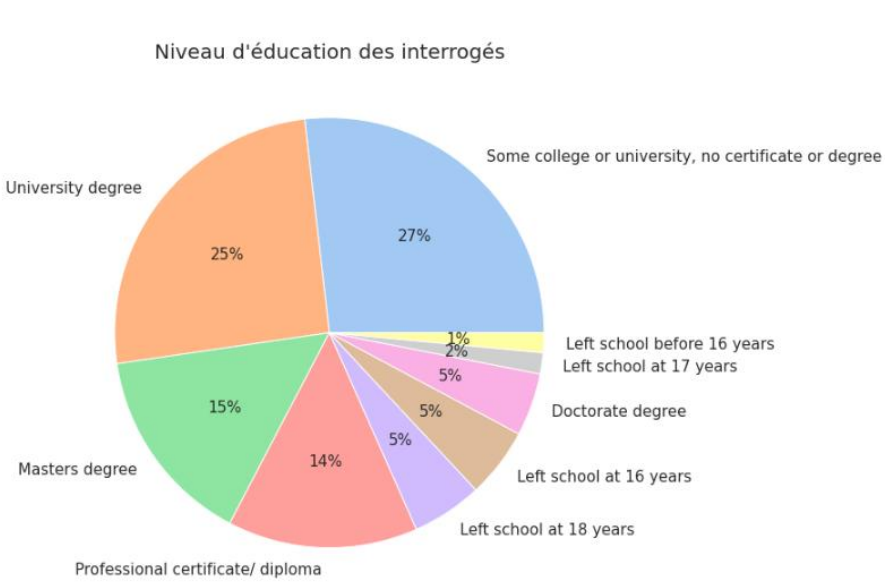


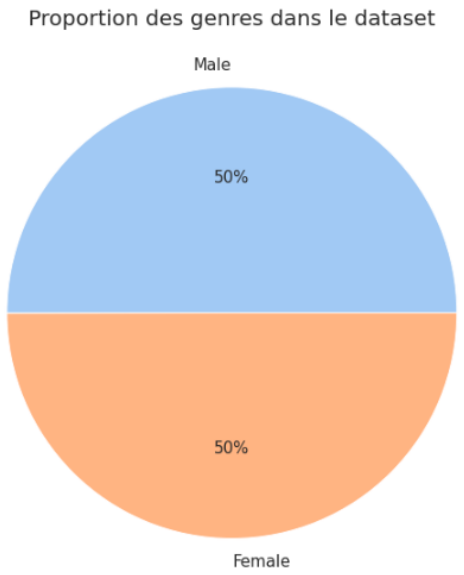
- Surreprésentation de l'ethnie « White » dans le dataset, et très peu de personnes âgés.

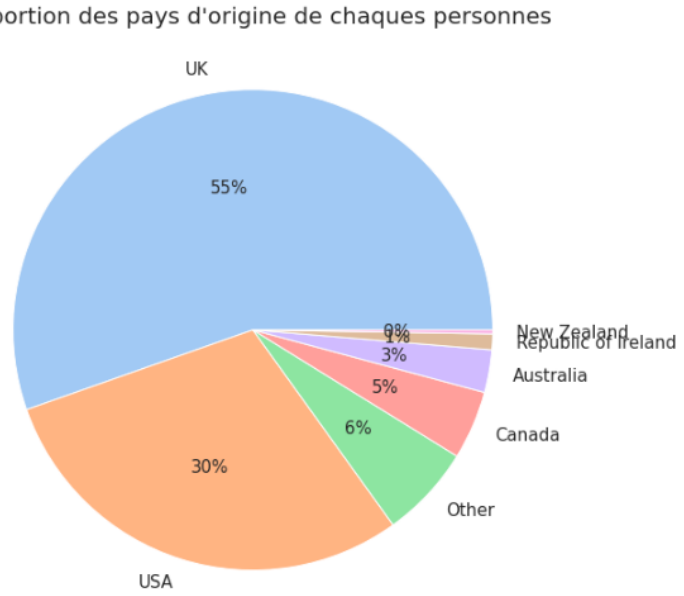
- Nous allons regrouper les personnes de plus de 55 ans et enlever la colonne « Ethnicity » pour rééquilibrer ces colonnes

Analyse du Dataset

Données personnelles







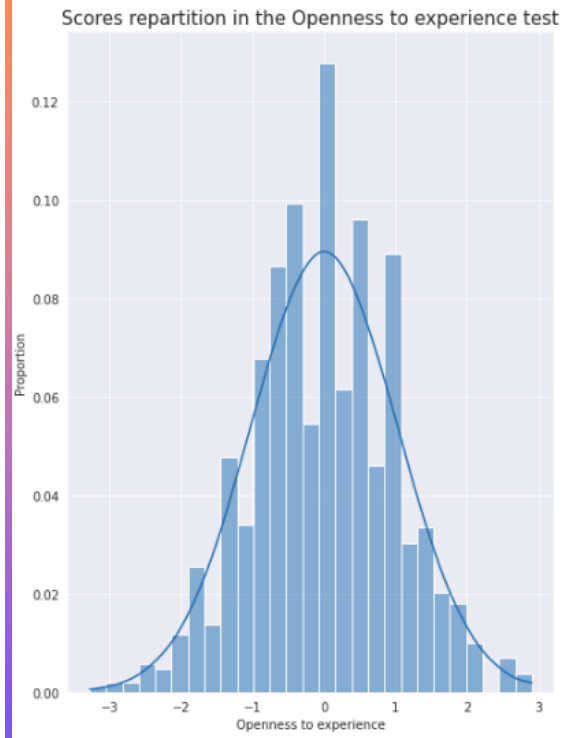
-Peu de personnes ayant quittés l'école avant 18 ans ainsi que d'Irlandais et de Néo-Zélandais

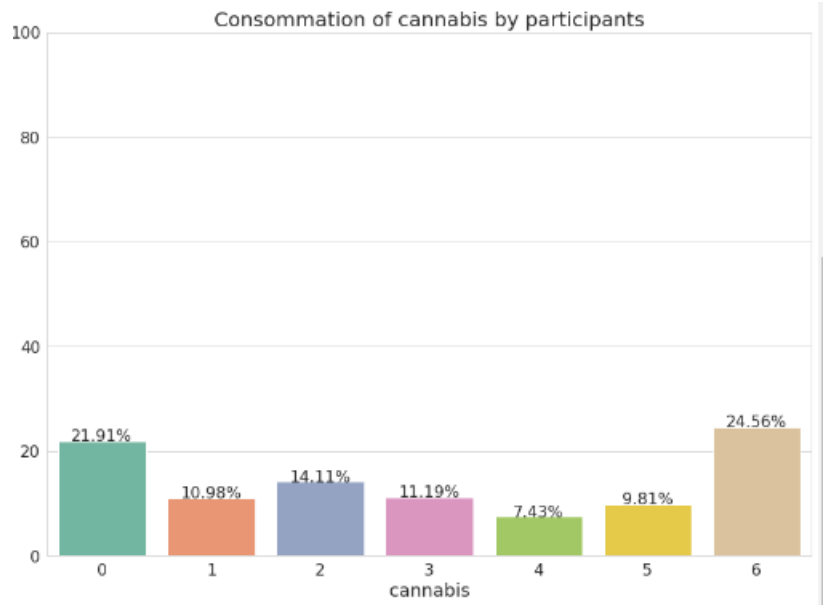
-Nous allons regrouper les gens ayant quittés l'école avec 18 ans.
On regroupe aussi les Irlandais avec les Britanniques et les Néo-Zélandais avec les Australien

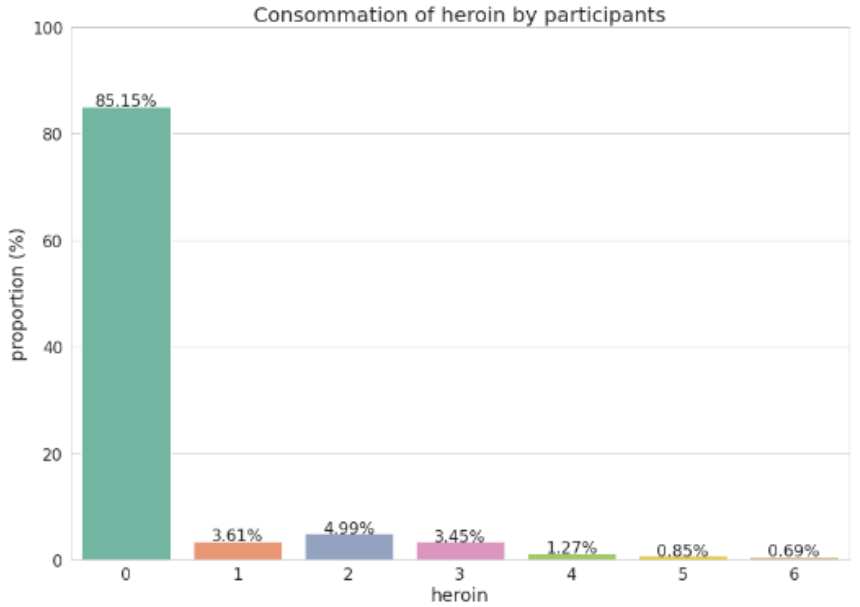
Analyse du Dataset

Drogues et tests de personnalités

- 0: jamais consommé
- 1: consommé il y a plus de 10 ans
- 2: durant la dernière decennie
- 3: cette année
- 4: ce mois
- 5: cette semaine
- 6: hier



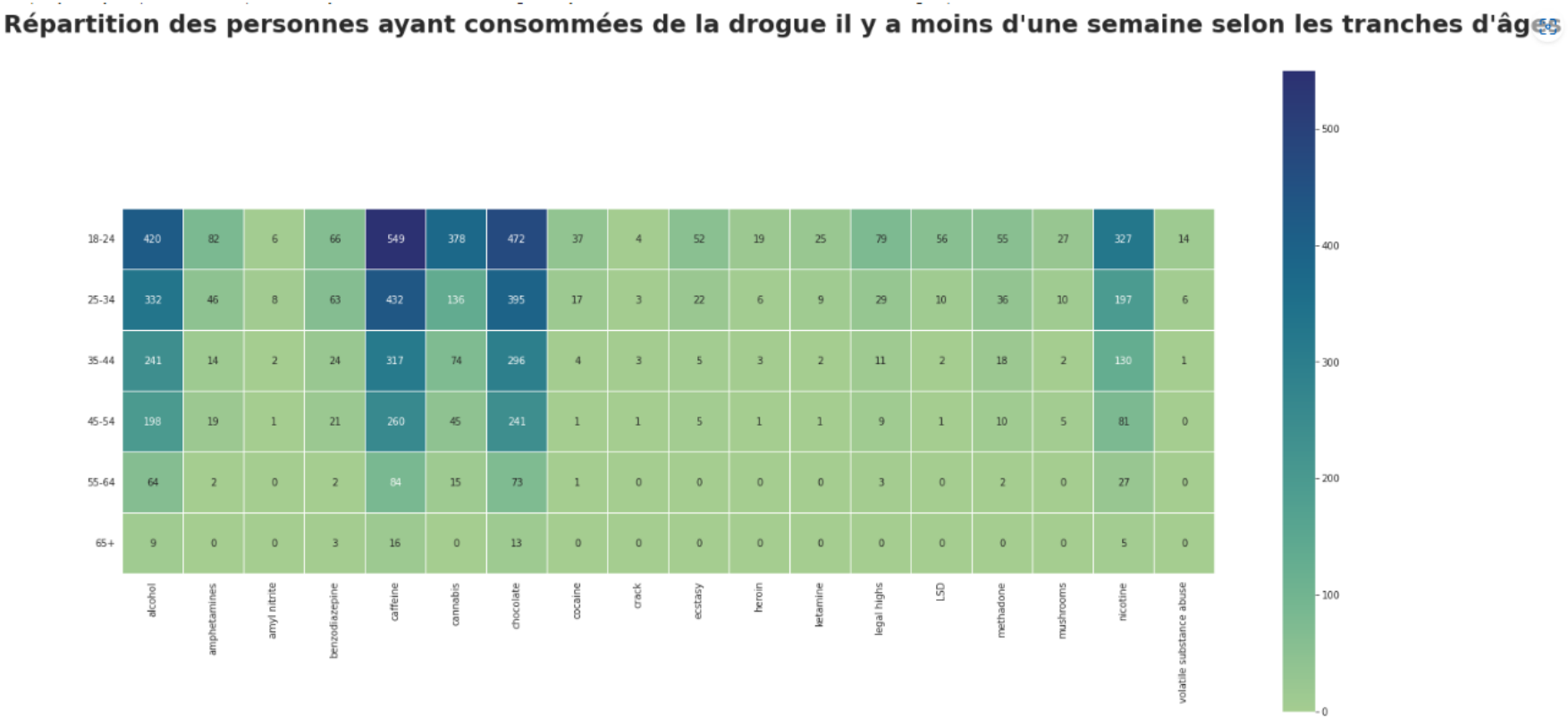




- Les scores des tests de personnalités suivent une loi normale
- On retrouve dans notre dataset des drogues légales (chocolat, caféine, nicotine et alcool) et illégales (héroïne, cannabis, LSD….)
- On ne prendra en compte que les drogues illégales sans le cannabis (légal dans certains états des USA et surreprésenté parmi les drogues illégales dans le dataset.). On considèrera comme « drogué » les personnes ayant consommé une des drogues illégales il y a moins d'une semaine (classe 5 et 6)

Analyse du Dataset

Heatmaps



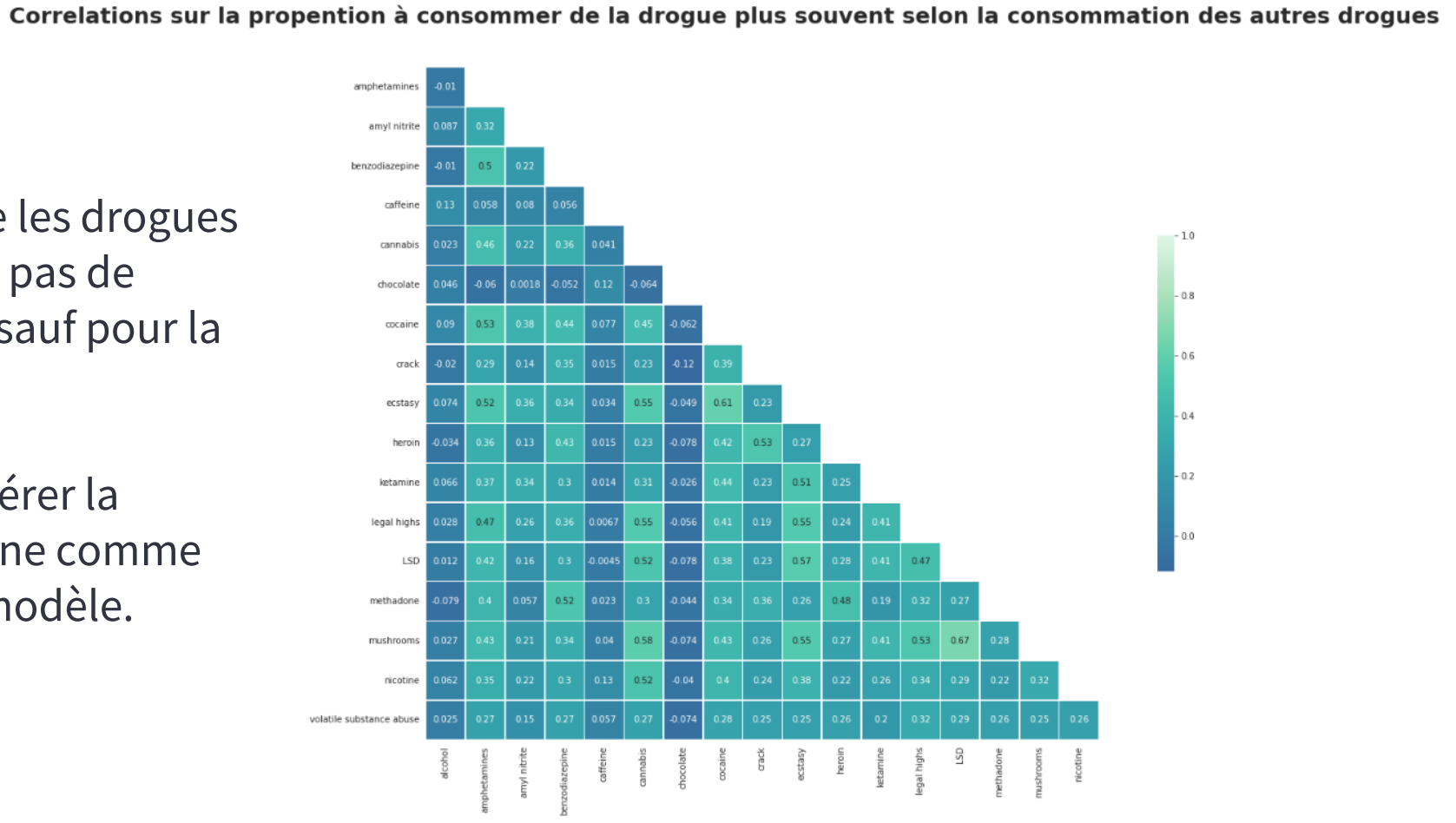
Trop peu de représentants dans certains croisements (vrai pour toutes des données, pas que pour l'âge), d'où la nécessité de regrouper les classes (features et target) pour éviter que le modèle overfit.

Analyse du Dataset

Matrices de corrélations

- Aucune corrélation entre les drogues illégales et légales (donc pas de logique à les regrouper) sauf pour la nicotine.

- Nous allons donc considérer la consommation de nicotine comme une feature pour notre modèle.



Création du modèle

Preprocessing

	Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	Impulsiveness	Sensation seeking	is_Australia_NZ	is_Canada	is_UK_Ireland	is_USA	Gender	Age	Education	nicotine
0	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1.18084	0.0	0.0	1.0	0.0	0.0	0.5	0.0	-0.2
1	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575	0.0	0.0	1.0	0.0	1.0	0.0	1.5	0.2
2	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0.40148	0.0	0.0	1.0	0.0	1.0	0.5	0.0	-0.6
3	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084	0.0	0.0	1.0	0.0	0.0	-0.5	1.0	-0.2
4	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	-0.21575	0.0	0.0	1.0	0.0	0.0	0.5	1.5	-0.2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1762	-1.19430	1.74091	1.88511	0.76096	-1.13788	0.88113	1.92173	0.0	0.0	0.0	1.0	0.0	-0.5	-0.5	-0.6
1763	-0.24649	1.74091	0.58331	0.76096	-1.51840	0.88113	0.76540	0.0	0.0	0.0	1.0	1.0	-0.5	-0.5	0.4
1764	1.13281	-1.37639	-1.27553	-1.77200	-1.38502	0.52975	-0.52593	0.0	0.0	0.0	1.0	0.0	0.0	0.5	0.6
1765	0.91093	-1.92173	0.29338	-1.62090	-2.57309	1.29221	1.22470	0.0	0.0	0.0	1.0	0.0	-0.5	-0.5	0.2
1766	-0.46725	2.12700	1.65653	1.11406	0.41594	0.88113	1.22470	0.0	0.0	1.0	0.0	1.0	-0.5	-0.5	0.6

- On applique les modifications suggérées précédemment au dataset. On crée la target en donnant la valeur 1 aux personnes ayant consommées une drogue illégale (sauf le cannabis) il y a moins d'une semaine, et 0 pour les autres

- On applique un OneHotEncoding sur la colonne des pays, un Ordinal encoder sur l'éducation, la nicotine et le genre, puis on applique un RobustScaler sur l'âge, l'éducation et la nicotine.

Création du modèle

Méthodologie

- Nous avons testé six modèles de machine learning (SVC,KNN, régression logistique, ADAboost, RandomForest et XGBoost) en modifiant ou non les hyperparamètres.

- Nous avons évalué les modèles avec 4 métriques: le recall, la précision, le f1-score et le score ROC AUC, ainsi qu'avec des matrices de confusion et des courbes ROC.

- Nous utilisons des courbes précisions/rappel pour atteindre la valeur de recall souhaité (80%) en modifiant le seuil de probabilité du modèle qui lui permet de choisir si une personne est dans la catégorie 1 ou 0.

- Nous jugeons alors les modèles sur leur précisions.

Création du modèle

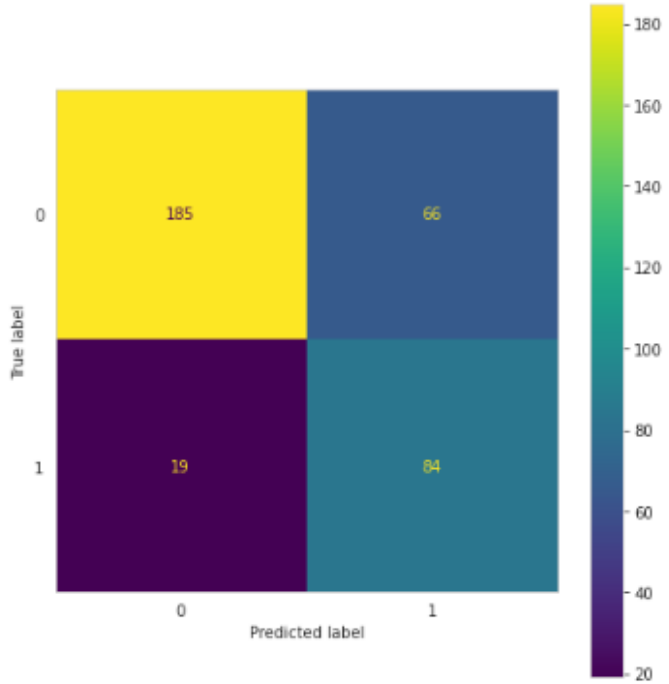
Résultats

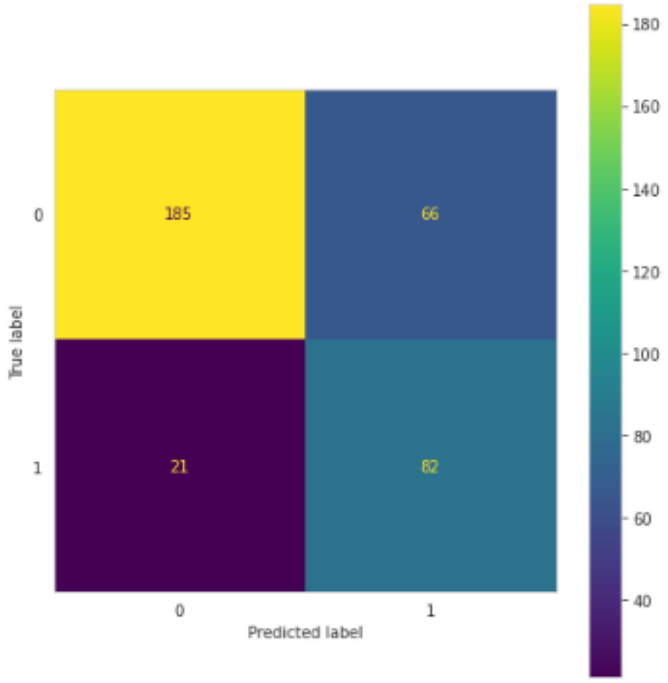
```
SVC:

f1 score= 0.6640316205533596
precision= 0.56
recall= 0.8155339805825242

RandomForest:

f1 score= 0.6533864541832669
precision= 0.5540540540540541
recall= 0.7961165048543689
```





Les deux meilleurs modèles parmi ceux testés sont le SVC et le random forest (avec tunning). On arrive avec ces deux modèles à une précision de 55% environ pour un recall de 80%.

Concrètement, on peut espérer que 80% des personnes « toxicomanes » soient prédites correctement si on assume que 45% des personne prédites comme toxicomane ne le sont en fait pas. Il faut pour cela que les seuils de probabilités soient de 23% (SVM) et 30% (RandomForest).

Toutefois l'utilisateur reste libre de se fier à son jugement, en choisissant comment classer un individu en regardant les probabilités prédites par le modèle.

Limite de l’analyse et du modèle

- Nos résultats ne s’appliquent qu’aux habitants des pays anglo-saxons

- Nombre relativement faible de donnée (1886 personnes interrogées, dont 8 menteurs affirmant avoir consommé la drogue fictive). Manque de représentativité des personnes âgées et de certains pays.

-Données légèrement biaisées, les personnes de l’échantillon consomment plus de drogue que la population globale. (selon le papier de recherche qui a utilisé les données)

- Nous devons faire attention à l’interprétation de la target: une personne ayant consommé une drogue illégale la semaine dernière n’est pas nécessairement toxicomane. Il faut prendre des précautions quant aux prédictions du modèle. On peut dire qu’une personne prédite positivement a des dispositions à consommer de la drogue, ce qui peut être une information intéressante pour l’organisme d’assistance sociale où nous travaillons.