Consommation de drogues, personnalité et données sociologiques

Souvigny Camille[*], Maliach-Auguste Romain[†], Barthélémy Jean-Gabriel[‡]

printemps 2021

Résumé

Ce document rend compte de l'analyse que nous avons fait du *Drug consumption (quantified) Data Set*, qui met en relation, pour 1885 participants et participantes à une étude, des fréquences de consommation de 18 drogues avec des résultats de tests de personnalité ainsi que des classes sociologiques (tranche d'âge, niveau d'études...). Le problème que nous avons tenté de résoudre est de créer un classifieur qui prédit le « type » de consommation de drogue (en fréquence et en « type » de drogue), en fonction des autres variables.

Pour cela, nous avons simplifié le problème. D'une part, nous avons transformé l'espace des fréquences de consommation, par une binarisation et une classification supervisée. D'autre part, nous avons transformé l'espace de la personnalité et des classes sociologiques (ci-après : espace des variables « explicatives ») par une analyse en composantes principales. Ensuite, nous avons essayé plusieurs algorithmes de classification.

1 Le problème

Le jeu de données (présenté ci-dessous) était accompagné de suggestions de problèmes, dont celui que nous avons choisi : prédire la consommation de drogue à partir des test de personnalité et des données sociologiques.

Nous avons simplifié ce problème de plusieurs façons. Au début, nous voulions prédire une fréquence de consommation (homogène aux fréquences du jeu de données), mais cela demandait de prédire 15 variables, chacune dans un espace discret à 7 modalités. Nous avons simplifié ce problème en nous ramenant à 2 modalités (« consommée récemment », ou : « consommée jamais ou il y a plus de dix ans »). Par ailleurs, nous avons regroupé les drogues (par classification non supervisée) en des classes où les fréquences de consommation sont homogènes.

Le résultat est un classifieur prédisant si l'utilisateur appartient ou non à une classe de consommation de drogues. Si un individu appartient à un groupe, cela signifie qu'on estime qu'il a, dans les dix dernières années, consommé une ou plusieurs des drogues des drogues de ce groupe.

2 Le jeu de données et nos prétraitements

Le jeu de données *Drug consumption (quantified)* est issu d'une étude d'Elaine Fehrman [1], Vincent Egan [2] et Evgeny M. Mirkes, [3] avec la participation de volontaires de l'hôpital Rampton de Retford, dans le Nottinghamshire. Il est distribué gratuitement sur un site de l'université de Californie à Irvine, « *UCI Machine Learning Repository* » [1].

Il s'agit d'un tableau dont les 1885 lignes sont des individus, et dont les colonnes sont des variables représentant des mesures hétérogènes.

Ces variables peuvent être partitionnées en trois classes : les variables mesurant les fréquences de consommation de chaque drogue, celles issues de deux tests de personnalité (NEO-FFI-R [7], et BIS-11 - ImpSS [2],[6]), et d'autres variables plaçant les individus dans des classes sociologiques : tranche d'âge, genre, niveau d'études, pays de résidence, et *ethnicity*.

En ce qui concerne les données de personnalité, nous ne pouvons pas discuter de la qualité scientifique des méthodes ayant produit ces variables : il faudrait une expertise scientifique ou médicale. En revanche, nous sommes en mesure de critiquer la qualité de certaines autres variables, ce que nous ferons ci-après. Puis, indépendamment de la qualité des données, nous justifierons notre raisonnement et nos calculs. Nos résultats auront une qualité inconnue, mais nous espérons que notre démarche est valable.

[*] camille.souvigny@etu.utc.fr
[†] romain.maliach-auguste@etu.utc.fr
[‡] jean-gabriel.barthelemy@etu.utc.fr

1. Elaine.Fehrman@nottshc.nhs.uk
2. Vincent.Egan@nottingham.ac.uk
3. em322@le.ac.uk

Critique des données

A priori, on ne possède aucune information sur la qualité de l'échantillon. On sait seulement que tous ces individus sont volontaires et viennent tous du même hôpital. On pourrait supposer qu'avec un si grand nombre de participants et la présence de chercheurs, la sélection des participants a été menée rigoureusement (sélection aléatoire ou redressée...) et que cet échantillon est représentatif de la population de l'hôpital. Mais, dans la notice accompagnant le jeu de donnée, nous n'avons aucune information permettant de dire dans quelle mesure les résultats que nous obtenons pourraient être généralisés à une plus grande population.

Cependant, des statistiques descriptives présentées en section 3 rendent compte de l'existence de problèmes de représentativité de l'échantillon. Sans informer sur la qualité de la méthode ayant produit cet échantillon, elles permettent de fixer des « bornes supérieures » à la généralisation qu'on pourrait faire de nos résultats. Par exemple, il n'y a que 18 individus ayant plus de 65 ans. Sans argument externe à l'analyse de ces données (expertise en médecine, statistiques de bonne qualité sur la population mondiale...), nous ne saurions affirmer que ces individus sont représentatifs et que nos classifieurs seront suffisamment robustes pour faire des prédictions sur un nouvel individu de cette classe.

Nous n'avons pas réalisé de pré-traitement corrigeant ce problème, puisque nous n'en connaissons pas l'étendue.

La façon dont les fréquences sont exprimées (les mêmes 7 classes pour toutes les drogues) fait que certaines variables apportent peu d'information. Par exemple : la consommation de chocolat est très concentrée dans les « hautes » fréquences, donc presque tous les individus sont dans seulement une ou deux classes. Pour trois drogues (café, chocolat, alcool), la corrélation entre variables explicatives et fréquences de consommation est nulle ou quasi-nulle, et nous ne pouvons pas dire si cela est intrinsèque aux drogues (ce seraient des drogues universelles, que tout le monde consomme de la même façon ?) ou si cela est simplement dû à un sous-échantillonage (la définition de classes de fréquence plus fines aurait pu permettre de mesurer des modalités différentes selon les individus).

Certaines mesures pourraient être plus fiables que d'autre. Par exemple, on pourrait imaginer qu'en clinique, il est soit aisé pour le personnel médical de déterminer (approximativement) l'âge, ce qui pourrait permettre de contrôler les données entrées par les participants. (Ce n'est qu'une spéculation, nous n'avons pas d'information sur la qualité des mesures). De même, les tests de personnalité ont été réalisés en clinique, par des professionnels, donc on pourrait espérer qu'ils soient de bonne qualité.

En revanche, on sait que les fréquences de consommation n'ont pas été mesurées par des scientifiques dans la nature, mais ce sont les sujets qui (de leur plein gré) ont fourni ces données en remplissant des questionnaires. On peut douter de leur objectivité dans la mesure où ils produisent les données qui les concernent. Il existe peut-être des pré-traitements permettant de corriger ces problèmes, et nous en avons appliqué un qui était prévu par la notice du jeu de données.

Afin de réduire un biais de « surestimation », Fehrman *et. al.* ont ajouté, dans leur questionnaire, une drogue fictive (la *Semeron*), et ont demandé aux volontaires de communiquer leur fréquence de consommation pour cette drogue. Nous avons choisi de supprimer de notre jeu de donnée les individus ayant répondu qu'ils avaient consommé de la *Semeron*. Ce n'est pas une correction parfaite dans la mesure où certains individus pourraient avoir prétendu avoir consommé de la Semeron pour une raison autre que du mensonge et de la surestimation de leur consommation (par exemple : mauvaise mémoire à long terme, erreur de saisie...). Pour l'exemple de l'erreur de saisie, on voit bien qu'en ne supprimant que les individus ayant commis une erreur de saisie sur la semeron (et pas les individus ayant commis une erreur de saisie sur d'autres variables), on ne résoud pas complètement le problème. Il aurait été intéressant de mesurer la « surestimation » d'une façon plus fiable (par exemple, avec un nombre de questions-témoin plus élevé, mais présentées d'une façon qui ne fassent pas comprendre à l'individu qu'il s'agit de témoins).

Au final, nos pré-traitements sont les suivants :

- suppression des individus prétendant avoir consommé de la *Semeron*
- filtre passe-bas : suppression des drogues dont la distribution des fréquences est concentrée dans les valeurs élevées, et qui sont peu corrélées aux variables de personnalité/sociologie : Alcool, Café, Chocolat
- suppression d'une variable illégale (« Ethnicity »). [4]

4. Même si ce traitement avait été autorisé par la CNIL, l'éthique aurait probablement interdit de fabriquer une machine qui prend une décision en se basant sur l'*Ethnicity* d'une personne. (Mais peut-être qu'il serait moralement plus acceptable de faire des travaux ayant pour objectif d'inciter la puissance publique à cibler ses efforts de prévention en montrant que certaines classes sociologiques ont des modes de consommation différents ? Cela dit, même des statistiques agrégées sur les individus, et descriptives, pourraient être utilisées contre des classes sociologiques, pour renforcer des discours incitant à les discriminer.)

3 Statistiques descriptives préliminaires

Des statistiques descriptives sont déjà fournies, variable par variable, dans la notice accompagnant le jeu de données (sur le site de l'UCI cité en section 2), ainsi que dans la publication [3] accompagnant le jeu de données. Nous avons choisi de ne pas refaire les mêmes analyses exploratoires, mais de les utiliser, et de suivre les pistes suggérées par Fehrman *et. al.*.

Cependant, nous allons rapidement paraphraser leurs résultats ici, en sélectionnant les plus intéressants.

Les individus de plus de 54 ans sont sous-représentés dans l'échantillon.

Les individus vivant dans des pays non anglophones ou en nouvelle-Zélande sont extrêmement sous-représentés.

Les résultats aux tests de personnalité semblent tous suivre des lois normales. Par exemple, la figure 1 représente la distribution des *n-scores*. Il serait intéressant de comparer les paramètres estimés de ces lois avec ceux mesurés par ailleurs sur d'autres populations. Une expertise scientifique ou médicale permettrait de décider si ces distributions sont « normales », et donc si les résultats obtenus avec notre échantillon sont généralisables.

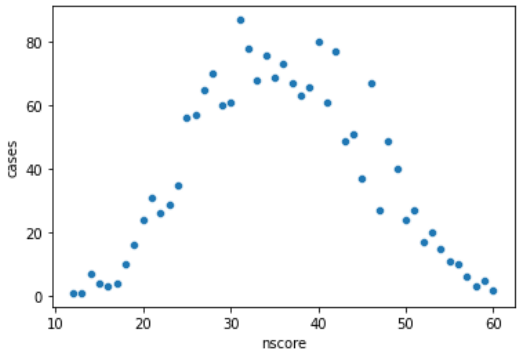

Figure 1 – Distribution des n-scores.

4 Transformation de l'espace des fréquences

4.1 Fréquences, modalités, binarisation

L'information que nous avons sur la consommation de drogues est, pour chaque individu, la fréquence à laquelle il consomme chaque drogue. Plus précisément, il ne s'agit pas tout à fait d'une fréquence, mais de la différence entre la date courante et la date de la dernière consommation. (Si la différence est grande alors la fréquence est faible ; si la différence est faible, alors la fréquence est peu connue). [5] Nous continuerons à utiliser le terme impropre de « fréquence » à la place de « inverse de la différence entre la date courante et la date de la dernière consommation ». Cette utilisation est abusive lorsqu'on parle de « fréquence » élevée : si la différence est petite, alors la fréquence n'est pas bien connue (puisqu'elle n'est connue qu'à l'intérieur d'un petit intervalle de temps). Mais cet abus permet plus de concision.

La « fréquence » est mesurée par une variable à sept modalités :

- la drogue n'a jamais été consommée
- consommée il y a plus de dix ans
- consommée au cours de la dernière décennie
- consommée l'année dernière
- consommée le mois dernier
- consommée la semaine dernière
- consommée hier.

Afin de simplifier le problème, nous avons réduit le nombre de modalités : la classe « au cours de la dernière décennie », et la classe « jamais ou il y a plus de dix ans ».

Cette binarisation était suggérée dans la notice accompagnant le jeu de données. Cependant, on peut critiquer l'arbitraire de cette frontière de dix ans, notamment si on souhaite utiliser ces résultats pour prédire une consommation future. Sans aucune recherche documentaire ni rigueur scientifique, et seulement pour illustrer l'arbitraire de cette frontière, nous avançons que l'évolution au cours du temps de la probabilité de « rechute » est très variable d'une drogue à une autre. À présent, nous formulons la même critique sans recourir à un argument mal fondé : on pourrait vouloir placer cette frontière plus ou moins loin, pour que l'assimilation des non-consommateurs aux consommateurs anciens aie plus ou moins de sens, et ce sens pourrait dépendre de chaque variable ainsi que de l'utilisation qu'on veut faire du classifieur final [6]. Selon le contexte, on pourrait par exemple vouloir modéliser le temps différemment, par exemple l'exprimer de façon relative à l'âge, ou encore créer des classes de fréquence différentes selon les âges, les drogues, ou d'autres variables. Voici un nouvel exemple sans fondement scientifique mais qui

5. mais dans ce document, nous ne ferons pas cette distinction sémantique, puisque nous ne chercherons pas (directement) à tirer des conséquences biologiques ou financières qui dépendraient d'une fréquence ou d'une date de dernière consommation

6. Souhaite-t-on étudier la légalité, de nocivité, un probabilité de rechute ?... Ces notions pourraient de la drogue, de l'individu, du pays...

illustre notre propos : on pourrait se dire que pour les individus jeunes, de 20 ans, les consommations sont nécessairement contenues dans les hautes « fréquences », et que les « dix dernières années » représentent une partie très importante de leur vie. Ainsi, une consommation ponctuelle et n'ayant pas d'impact sur le reste de la vie de l'individu pourrait être interprétée à tort comme une addiction, renforçant la corrélation entre l'âge et la consommation. Ainsi, selon la sémantique adoptée, qui dépend en fait de l'usage qu'on veut faire du classifieur, on peut critiquer la façon dont la consommation est exprimée.

Sans beaucoup de fondements scientifiques, mais seulement afin de simplifier le problème, nous proposerons une binarisation différente pour les « nicotinoïdes », en section 8.

4.2 Classification non supervisée dans l'espace des fréquences binaires

4.2.1 Critique et justification

Plusieurs travaux cités dans la notice, dont [3], suggèrent qu'il existent des « pléïades » de drogues étant sujettes à des modes de consommation similaires et provoquant des effets similaires. Nous nous proposons, pour rendre le problème de classification plus simple (réduction du nombre de variables à prédire), de faire une classification non supervisée des drogues en fonction de la façon dont elles sont consommées (en fréquence).

Nous ne recourrons pas à des arguments chimiques ni éthologiques pour affirmer qu'on peut assimiler certaines drogues les unes aux autres. Par ailleurs, une faiblesse de cette approche est que la similarité des modes de consommation pourrait être expliquée par des propriétés non intrinsèques aux drogues (chimie, biologie, toxicologie...) mais par des propriétés extrinsèques (prix, disponibilités, distribution géographique, incitation ou réprobation dans un milieu social donné...). Or, il ne faut pas oublier que l'objectif de notre travail est de créer un classifieur qui prédit la consommation en fonction de mesures sur l'individu : ces facteurs extrinsèques, qui génèrent notre partition, pourraient se retrouver dans les variables explicatives (par exemple : le niveau d'études déterminerait le milieu social, qui déterminerait la pression sociale, qui déterminerait les modes de consommation).

Tant que l'on n'a pas clairement déterminé l'objectif, l'usage qui sera fait du classifieur qu'on cherche à construire, on ne peut pas déterminer si c'est une bonne chose qu'un certain phénomène influence à la fois les variables explicatives et la génération des partitions (donc, dans une certaine mesure, il influence les variables expliquées). Dans certains cas d'usage, on pourrait vouloir mettre en évidence une forte corrélation ; dans d'autre cas, cette corrélation pourrait être vue comme artificielle.

Nous allons illustrer ce problème en gardant l'exemple (inventé et non fondé sur de vrais travaux documentaires ni scientifiques) où le milieu social détermine les modes de consommation des drogues. Lorsqu'on suppose que l'utilisateur de notre classifieur souhaite étudier les modes de consommation des drogues parce qu'il s'intéresse à leurs effets biologiques (par exemple : l'addiction), alors il s'intéresse bien à des propriétés intrinsèques aux drogues, qui sont constatées cliniquement, et expliquées par la biologie. Par conséquent, il lui convient d'obtenir des résultats où les drogues ayant des modes de consommation similaires sont assimilées les unes aux autres dans un cluster (non distinguées). Or, si on suppose que notre classification non supervisée est réalisée sur un échantillon tiré au hasard et représentatif d'une certaine population dans laquelle il existe plusieurs milieux sociaux qui déterminent les modes de consommation de drogue, alors les clusters qu'on obtient sont des clusters de drogues qui sont consommées de la même façon pas seulement pour des raisons chimiques, mais aussi pour des raisons « sociales ». On obtient donc des résultats difficilement exploitables pour cet utilisateur.

En revanche, la puissance publique ayant une approche sociologique, souhaitant rendre sa politique de prévention de la consommation plus ciblée, pourrait vouloir mettre en évidence qu'un groupe social homogène [7] a des consommations homogènes, sans avoir besoin de savoir si les drogues d'une même classe sont intrinsèquement similaires. (C'est-à-dire que les mécanismes créant ces modes de consommation peuvent rester des boîtes noires.) [8].

Au final, on peut avoir de bonnes ou de mauvaises raisons de créer des clusters de drogues ; la notre est simplement de rendre le problème plus simple, en suivant une démarche suggérée par la notice du jeu de donnée.

7. remarquez que ce n'est plus un problème de prédiction individu -> classe de drogues, mais une « prédiction » classe d'individus -> classe de drogues. Mais cela montre tout de même qu'on peut souhaiter avoir une partition formant l'espace d'arrivée dont la génération est corrélée à une des variables explicatives.

8. Cela dit, la puissance publique pourrait vouloir cibler un groupe sociologique, mais aussi cibler des drogues, en fonction de leur nocivité. Une solution à ce problème pourrait être de pondérer les drogues par leur nocivité avant de faire la classification non supervisée. Une autre solution serait d'introduire la variable de nocivité dans l'espace à expliquer.

4.2.2 Mise en place de la classification non supervisée

Après binarisation, nous obtenons un espace où la distance la plus adéquate est celle de Hamming, qui compte le nombre de bits qu'il faut modifier pour passer d'un point à un autre. La figure 2 illustre le calcul de cette distance pour un espace à trois variables binaires.

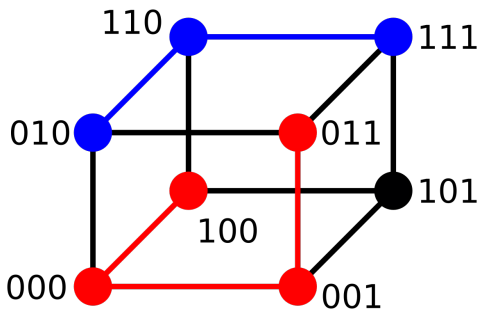

Figure 2 – Illustration du calcul de la distance de Hamming pour trois variables binaires. Colin M.L. Burnett, ©GFDL

Dans un tel espace, on ne peut pas appliquer directement des algorithmes de classification tels que les k-means, qui ont besoin de calculer des points moyens. Nous avons commencé par calculer une classification ascendante hiérarchique (CAH), à l'aide du module *AgglomerativeClustering* de *scikit-learn* [5]. Cette approche permet de ne pas devoir choisir un nombre de classes *a priori*. Cependant, il y a un paramètre que nous ne savions pas choisir : le critère de saut. Nous avons donc exécuté trois CAH, avec trois critères différents.

L'article de Fehrman *et. al.*[3] suggérait de former trois « pléïades » :

```
heroinPl =
["Crack", "Coke", "Meth", "Heroin"]
ecstasyPl =
["Amphet", "Cannabis", "Coke", "
    Ketamine", "LSD", "Mushrooms", "
    Legalh", "Ecstasy" ]
benzoPl = ["Meth", "Amphet", "Coke"]
```

L'ensemble de ces pléïades ne forme pas une partition des drogues, puisque certaines drogues appartiennent à plusieurs pléïades (par exemple, la cocaïne appartient aux trois). Cependant, on peut s'en inspirer.

Comme la classe est plus « forte » que la pléïade (une drogue ne peut appartenir qu'à une classe), on peut supposer qu'une bonne partition contiendra un nombre de classes supérieur au nombre de bonnes pléïades. Nous avons donc utilisé la CAH pour créer des partitions à 3, 4 et 5 classes, puis calculé les inerties [9] intra-classes pour comparer ces classifications. Plus précisément, nous avons utilisé la métrique scalaire dite de l'*inertia*, proposée par *sk-learn*[5], qui est la somme des distances au carré de chaque point au centr(oïd)e [10] de sa classe. On obtient un scalaire évaluant la qualité d'une partition.

Outre le nombre de classes, un autre paramètre que nous devions choisir est le critère de saut. Dans le doute, nous avons essayé les trois à notre disposition : saut minimal, maximal, moyen. Les sauts moyen et maximal ont donné exactement la même hiérarchie, et celle du saut minimal en différent seulement par un branchement, qui isole la Benzos des autres non-« nicotinoïdes » [11]. (On verra à la fin de cette section que le Benzos est un *outlier* (cas particulier) qui mérite d'avoir sa propre classe).

Par la suite, lorsque nous présenterons les prochains algorithmes utilisés pour la classification non supervisée, il sera implicite que s'il fallait choisir entre plusieurs paramètres (initialisation, heuristique, etc.), nous en avons essayé plusieurs proposés par *sk-learn* et conservé ceux qui donnaient les meilleurs résultats.

Une autre façon de former des partitions est l'algorithme des k-médoïdes. Il s'agit d'une « généralisation » de l'algorithme des k-moyennes aux espaces discrets, qui ne calcule pas des points moyens, mais utilise des centroïdes, qui sont les représentants les « plus centraux » (au sens d'une distance donnée) des nuages. L'algorithme des k-médoïdes possède beaucoup de propriétés communes à celui des k-moyennes, mais est plus robustes par rapport aux *outliers*.

Dans un espace discret et « plutôt rempli » (contenant un nombre de points « élevé » par rapport au nombre total de points possibles), l'initialisation est beaucoup plus déterminante qu'en k-moyennes. Dans notre cas particulier, il quasiment impossible d'obtenir une convergence lorsqu'on prenait une initialisation aléatoire. Dans le cas général, il faut aussi faire attention à éviter les minima locaux. Dans notre cas, nous savions à l'avance à quoi nos résultats devaient ressembler (puisque nous avions déjà les pléïades fournies par la notice du jeu), alors nous savions s'il était possible de faire mieux.

9. au sens de la distance de Hamming, puis au sens de la distance euclidienne dans l'espace décrit dans la section AFTD.
10. dans le cas où on utilise une distance de Hamming, le calcul utilise des centroïdes ; dans le cas où on utilise une distance euclidienne, il s'agit de centres.
11. Nous appelons nicotinoïdes (peut-être improprement) l'ensemble « Nicotine, Cannabis ».

5

À part les k-médoïdes et la CAH, il n'existe pas beaucoup de façons de former des classes homogènes dans un espace binaire. Nous avons donc effectué une AFTD qui nous a permis d'obtenir des espaces euclidiens à deux ou trois dimensions, avec un *stress*[12] d'environ 12%. Dans ces espaces, nous avons pu appliquer d'autres algorithmes de classification tels que les k-moyennes.

La plupart des partitions que nous avons obtenues ne différaient que de quelques drogues (1 ou 2). Certains algorithmes non déterministes « hésitaient » pour certaines drogues qui apparaissaient tantôt dans une classe, tantôt dans l'autre ; ces classes étaient d'ailleurs peu homogènes. Nous avons identifié ces éléments comme *outliers* : nous leur avons donné leur propre classe. Ce choix a été guidé par les inerties (retirer seulement l'outlier provoque une chute d'inertie beaucoup plus grande que retirer 3 ou 4 autres éléments du nuage), mais aussi par de la visualisation 2D et 3D de l'espace euclidien qui mettait bien en évidence que certaines drogues se trouvaient « à mi-chemin » de deux ou trois groupes de drogue, et qu'il était impossible de décider de quel groupe elles se rapprochaient le plus.

Notre meilleure partition a été obtenue automatiquement, puis corrigée « à la main » lorsqu'il a fallu isoler quelques outliers. Elle est présentée dans le tableau 1. Afin de rendre compte de tous nos essais, nous avons résumé par l'*inertia*, dans le tableau 2, la qualité de toutes les partitions que nous avons trouvé. L'*inertia* est mesurée par la distance de Hamming, et par la distance euclidienne dans l'espace transformé après AFTD. Il faut prendre en compte que l'AFTD a modifié les données avec un *stress* d'environ 12%, donc l'*inertia* euclidienne est une moins bonne mesure de la qualité des partitions. Nous nous en sommes tout de même servi car elle a l'avantage de correspondre aux visualisations 2D et 3D.

Table 1 – Notre partition des drogues (référencée sous le nom *alamano* dans le tableau 2). L'inertie est calculée avec la distance de Hamming, et avec la distance euclidienne dans l'espace obtenu après AFTD ($stress \approx 12\%$).

n° de classe	drogues	inertie Hamm.	inertie Eucl.
0	LSD, Mushrooms, Legalh	3	0,02
1	Nicotine, Cannabis	2	0,02
2	Coke, Ecstasy	2	0,01
3	Meth, VSA, Ketamine, Heroin, Crack, AMyl	6	0,09
4	Benzos	0	0
5	Amphet	0	0

12. au sens de sk-learn : « *sum of squared distance of the disparities and the distances for all constrained points* ». Ici, la distance de Hamming est vue comme une dissimilarité qu'il faut transformer en distance euclidienne.

Table 2 – *inertia* : mesure globale de la qualité des partitions, à partir d'inerties calculées dans l'espace de Hamming ou dans l'espace euclidien 3D obtenu après AFTD (stress ≈ 12%). Dans notre nomenclature, le nombre à la fin est le nombre de classes.

Partition (réf)	obtenue par	hamming	3D
cahcompl3	CAH saut moyen CAH saut max	15	3,39
cahmin3	CAH saut min	15	3,4
kmdrand3	k-médoïdes init. aléat.	N-A	N-A[15]
kmd++3	k-médoïdes++[13]		2,3
kmd++4	k-médoïdes++	13	2,46
kmd++5	k-médoïdes++	15	2,02
kmdhr3	k-médoïdes init. heur.[14]		2,5
kmdhr4	k-médoïdes init. heur.	15	2,02
AFTD3Dkm4	k-moyennes après AFTD 3D	15	3,15
alamano6	à la main	13	0,14

13 14 15

On notera que la partition formée à la main possède 1 ou 2 éléments de plus que les autres ; dans cette mesure, on pourrait se dire qu'il est « facile » d'obtenir de meilleures performances, et qu'il aurait suffi d'imposer l'obtention d'un plus grand nombre de classes. Cependant, pour certaines méthodes, cela est très difficile voire impossible. Pour la CAH, il est difficile de fixer le seuil où l'on sépare la hiérarchie en plus de trois classes (d'autant que les indices ne sont pas très déterministes). De même, lorsqu'on impose un plus grand nombre de classes à des algorithmes « agglomératifs » (k-moyennes, k-médoïdes), lorsqu'on répète la procédure, on obtient tout de même des « hésitations » pour les *outliers* (ils se retrouvent alternativement dans une classe ou dans une autre). Étant donné que le but était de simplifier le problème en réduisant le nombre de variables à prédire, nous n'avons pas spécialement intérêt à créer un grand nombre de classes. C'est pourquoi nous avons préféré isoler, d'une part, des *outliers*, et d'autre part, quelques clusters très remplis et très homogènes, plutôt qu'un grand nombre de clusters moins homogènes.

5 Analyse en composante principales

Une fois notre classification non supervisée des variables de consommation de drogue terminée, nous nous

13. Initialisation analogue à k-means++ : on commence par des points qui localement sont déjà des centroïdes.

14. Documentation sk-learn : « heuristic picks the n clusters points with the smallest sum distance to every other point. ».

15. Nous n'avons pas réussi à faire converger l'algorithme en imposant 3, 4 ou 5 classes.

6

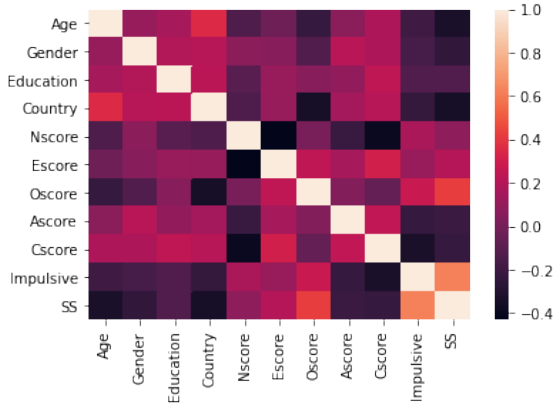

Figure 3 – Corrélations entre variables explicatives

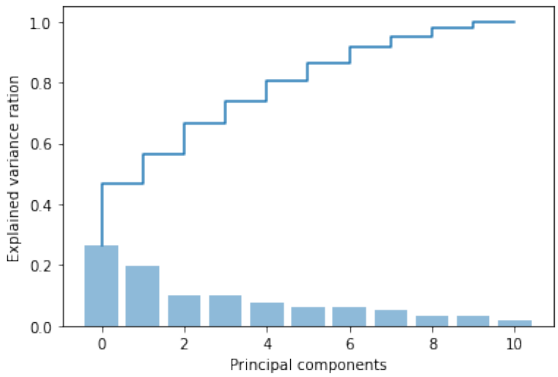

Figure 4 – Variance expliquée cumulée.

sommes intéressées aux onze variables explicatives :

- variables socio-économiques : *Age*, *Gender*, *Education*, *Country*,
- scores produits par les deux tests de personnalité : *Neuroticism*[16], *Extraversion*[16], *Openness to experience*[16], *Agreeableness*[16], *Conscientiousness*[16], *Impulsiveness*[17] et *Sensation Seeking*[17] (« SS »).

Afin de simplifier notre problème, nous souhaitions diminuer le nombre de variables explicatives. Certaines variables sont corrélées, c'est le cas des variables Impulsiveness et SS (corrélation supérieure à $0,7$), comme on peut le voir en figure 3.

Il est donc intéressant d'appliquer une méthode factorielle telle que l'analyse en composantes principales. En réalisant une ACP avec la bibliothèque *sk-learn*, nous avons pu réduire notre espace à 10 variables explicatives avec 98% de la variance expliquée et 9 colonnes pour 95% de la variance expliquée. La première a une variance expliqué de 24%, comme on peut le voir en figure 4.

6 Création des classifieurs

Suite aux différents traitements effectués sur la donnée, nous nous intéresserons dans cette partie à la construction des différents classifieurs. Nous avons testé de nombreuses techniques de classification : Adaboost, l'arbre décisionnel, la forêt d'arbres décisionnels, les K-plus proches voisins, le classifieur naïf bayésien, les réseaux de neurones et les séparateurs à vaste marge.

Nous avons pour chaque classifieur estimé l'erreur empirique en réalisant une moyenne des erreurs commises à partir d'une validation croisé avec proportionnalité des classes. C'est à dire, que chaque jeu de données contient la même quantité de personnes consommant ou non ce type de drogue. Les données étaient découpées en 10 part égales une seule étant utilisé pour mesuré l'erreur.

Chaque performance est à mettre en perspective avec le classifieur à vote majoritaire qui prédit la classe la plus présente dans la donnée : ainsi, on obtient un score à battre.

Au vu du peu de dimensions enlevé par l'ACP, nous avons testé les classifieurs sur l'ensemble des données, avec et sans ACP. Les performances de nos classifieurs sont légèrement plus élevée avec les données issues de l'ACP.

Le réseau de neurones a été construit selon les mêmes modalités que l'article de E.M. Liu [4]. Il est donc composé d'une couche intermédiaire avec 4 neurones et une couche de sortie. Cependant nous n'obtenons pas de résultats aussi bons que ce que semblait obtenir cet article. Cela est du au fait que l'article présente une classe {Héroïne, Ecstasy, Crack, Cocaïne} sans justification avec les mêmes critères de discrétisation que les nôtres, ainsi nous n'avons pas tester les réseaux de neurones pour cette classe.

7 Performances des classifieurs

Les performances sont présentées dans la table 3, par classifieur, par classe, et par nombre de dimensions. L'espace à 11 dimensions est l'espace donné directement par le jeu de données ; les espaces à 10 et 9 dimensions sont obtenus après une ACP, en retirant le dernier ou les deux derniers axes principaux. Nous avons utilisé de la validation croisée.

16. au sens de la méthode NEO-FFI-R [7]
17. Au sens des méthodes ImpSS [6] et BIS-11 [2].

7

Table 3 – Taux d'erreur des classifieurs pour les 4 classes, sans (dim = 11) et avec ACP

dim	Vote maj.	Adab. opt.	Tree	rand. forst.	knn opt.	Naive	SVC	réseau neur.
	Classe Coke, Ecstasy							
11	0.47	0.27	0.37	0.27	0.27	0.27	0.27	0.31
10	0.47	0.28	0.36	0.29	0.27	0.27	0.27	0.33
9	0.47	0.28	0.35	0.28	0.27	0.26	**0.26**	0.40
	classe LSD, Champignons, *Legal high*							
11	0.49	0.21	0.27	0.19	0.21	0.20	0.20	0.25
10	0.49	0.21	0.28	0.21	0.21	0.20	0.22	0.30
9	0.49	0.21	0.29	0.21	0.20	0.20	**0.19**	0.30
	classe Meth, VSA, Ketamine, Heroin, Crack, AMyl							
11	0.45	0.29	0.37	0.32	0.28	0.28	**0.27**	0.32
10	0.45	0.28	0.38	0.32	0.31	0.29	0.28	0.33
9	0.45	0.29	0.36	0.34	0.31	0.29	0.28	0.35
	classe Nicotine, Cannabis							
11	0.23	0.21	0.27	0.23	0.21	0.23	0.19	0.22
10	0.23	0.22	0.26	0.22	0.21	0.20	**0.18**	0.31
9	0.23	0.22	0.25	0.23	0.21	0.20	0.19	0.29

Pour les classifieurs Adaboost, KNN et Random Forest, nous avons optimisé les hyper-paramètres avec GridSearchCV.

Par exemple, le nombre optimal de voisins pour le classifieur de la classe $\{\text{Coke}, \text{Ecstasy}\}$ est 173. Les paramètres optimaux du classifieur RandomForest de $\{\text{Coke}, \text{Ecstasy}\}$ pour les données issues de l'ACP sont : 170 estimateurs avec le critère Gini.

Nous avons établi la liste des critères $n_{\text{estimators}}$ et $n_{\text{neighbors}}$ avec des générateurs logarithmiques de tableaux.

Les résultats sont médiocres, seule la classe LSD, Champignons, *Legal high* obtient, des résultats assez bon par rapport au vote majoritaire.

On observe que le classifieur le plus performant est le séparateur à vaste marge avec le noyau par défaut, une fonction de base radiale (RBF). Il serait intéressant d'explorer d'autres noyaux.

Les réseaux de neurones semblent mieux fonctionner sur les données brutes que sur les données ayant subi une ACP.

Dans le cas de la classe $\{\text{nicotine}, \text{cannabis}\}$, 77% des individus en avaient déjà consommé.[18] Les performances obtenues par nos classifieurs sont très mauvais au vue du classifieur à vote majoritaire. Cela nous pousse à revoir les seuils à partir des quels il faut considérer qu'un individu est consommateur d'une drogue. Nous verrons cela dans la partie suivante.

18. On pourrait imaginer beaucoup d'explications à ce seuil très élevé. Sans fondement scientifique, on pourrait par exemple avancer que ce sont des drogues plus accessibles, à la fois matériellement et moralement (drogues « douces »). Leur consommation est moins répréhensible.

8 Modification du critère de binarisation

8.1 Justification

Les proportions de consommation de chaque classe étant très différentes, particulièrement pour la classe nicotine-cannabis, nous nous sommes penchés sur une binarisation différente selon les classes de drogues. En effet, le cannabis et la nicotine sont des drogues beaucoup plus consommées que les autres drogues de cette études. Nous considérons par exemple que la consommation de nicotine dans le 10 dernières années n'est pas aussi significatives que la consommation d'une des autres drogues.

Nous avons alors attribué une nouvelle binarisation à ces deux variables. Les deux classes choisies sont « au cours de la dernière année » et « jamais, ou il y a plus d'un an ».

On pourrait vouloir choisir d'autres binarisations : cela dépend en fait de l'usage qu'on veut faire du classifieur. Si l'on veut prédire des rechutes après 20 ans, ou si l'on veut deviner l'état des poumons, on ne choisira peut-être pas le même seuil. Nous avons choisi cette autre binarisation en espérant obtenir de meilleures performances, mais ce choix devrait être mieux justifié.

8.2 Performances

Les résultats de cette binarisation permettent une amélioration des proportions de nos classes. Mais les performances des classifieurs ne sont pas fortement améliorés.

Autre que la discrétisation des données citée précédemment, une binarisation des drogues corrélée avec la fréquence d'utilisation (pour créer des classes en meilleure répartition) serait une piste d'amélioration (à condition que cela aie du sens par rapport au cas d'usage du classifieur).

9 Conclusion

Après une simplification du problème et en ayant testé plusieurs modèles, nous avons obtenu des classi-

fieurs donnant des performances médiocres voire correctes. Mais il ne faut pas trop s'en réjouir : nous avons fait des choix qui simplifient le problème mais qui auraient dû être faits par des spécialistes du domaine, en fonction du cas d'usage du classifieur. Par exemple la binarisation des fréquences de consommation. Par ailleurs, nous n'avons pas d'argument permettant de dire que l'échantillon est représentatif. Notre classifieur n'est donc pas directement « utilisable ».

9.1 Pistes d'amélioration

Afin d'être plus indépendant du cas d'usage, on pourrait résoudre un problème moins simplifié, par exemple en abandonnant la binarisation. On pourrait aussi abandonner la classification non supervisée, et faire cette prédiction de consommation drogue par drogue. D'ailleurs, pour justifier nos simplifications, on pourrait essayer de résoudre le problème naïf, et vérifier que les performances sont moins bonnes.

Par ailleurs, nous avons montré à plusieurs reprises que la façon dont les fréquences était exprimées générait des effets de seuil susceptibles d'être mal interprétés. Selon le cas d'usage du classifieur, on pourrait exprimer le temps sous une forme relative à l'âge, ou encore le linéariser sous la forme de probabilités [19]. On pourrait aussi s'autoriser à prédire des probabilités d'appartenance à chacune de ces classes (logique floue).

Si on voulait conserver les classes de drogue, il serait intéressant de faire des statistiques descriptives à l'intérieur de chacune de ces classes, afin de mieux interpréter ce que signifie l'appartenance (prédite ou constatée) à l'une de ces classes. Par exemple, on pourrait vérifier, variable par variable ou après ACP, si certains individus sont sur-représentés ou sous-représentés dans chaque classe de consommation. Par ailleurs, on pourrait aussi vérifier si certaines drogues sont plus consommées que d'autres à l'intérieur d'une classe. On pourrait aussi leur associer une expertise en chimie, biologie, médecine ou toxicologie, afin de savoir si la partition a du sens. En effet, nous avons formé nos classes, nous avons fait en sorte que les fréquences de consommation soient homogènes, mais la chimie ou la nocivité des drogues ne l'est pas nécessairement. On ne pourra donc pas forcément utiliser notre classificateur pour faire des prédictions sur la santé, il faudrait d'abord une expertise en chimie, médecine ou toxicologie pour vérifier si les effets des drogues des classes que nous avons formées sont similaires. Pour le moment, nous prédisons, pour un individu, non seulement les drogues qu'il consomme, mais aussi la façon dont il consomme (en temps).

Et c'est pour cela qu'on pourrait aussi vouloir relâcher la contrainte d'une partition et revenir aux pléïades. En effet, en faisant des classes, nous avons supposés que, pour chaque drogue ou classe de drogue, il existait un seul type de consommation. En revenant aux pléïades, on admet qu'une même drogue est susceptible d'être consommée de différentes façons, d'appartenir à différentes classes de fréquence de consommation des drogues.

19. on transforme la modalité « consommation dans les dix dernières années » en une probabilité $p = 1/10$ de consommer de la drogue au cours de l'année prochaine.

Références

[1] Dheeru Dua et Casey Graff. *UCI Machine Learning Repository - Drug Consumption (quantified)*. 2017. URL : https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29.

[2] Héloïse Dupont et Jean Cottraux. *Evaluation dimensionnelle de l'impulsivité dans le trouble obsessionnel-compulsif*. Partie protocole expérimental, section 2.2.2, page 93. Université Lumière Lyon 2, 2002. URL : http://bibelec.univ-lyon2.fr/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,url,uid&db=cat07741a&AN=spb.245364&lang=fr&site=eds-live.

[3] E. Fehrman et al. *The Five Factor Model of personality and evaluation of drug consumption risk*. 2017. arXiv : 1506.06297 [stat.AP].

[4] Eric Max Liu. « Predicting illicit drug use with artificial neural network. » In : *European Journal of Humanities and Social Sciences* 3 (2019), p. 131-137.

[5] F. Pedregosa et al. « Scikit-learn : Machine Learning in Python ». In : *Journal of Machine Learning Research* 12 (2011), p. 2825-2830.

[6] Wikipedia contributors. *Alternative five model of personality — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Alternative_five_model_of_personality&oldid=1010890568. [Online ; accessed 30-May-2021]. 2021.

[7] Wikipedia contributors. *Revised NEO Personality Inventory — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Revised_NEO_Personality_Inventory&oldid=1015494897. [Online ; accessed 30-May-2021]. 2021.

Annexe

Table 4 – Paramètres optimisés des différents classifieurs

Algorithme	Classe	$n_{neighbors}$	$n_{estimators}$	critère
KNN	LSD,Mushrooms,Legalh	32		
	Nicotine, Cannabis	30		
	Met, VSA, K, H, Crack, AMyl	11		
	Coke, Ectasy	173		
Adaboost	LSD, Mushrooms, Legalh		113	
	Nicotine, Cannabis		7	
	Meth, VSA, Ketamine, Heroin, Crack, AMyl		22	
	Coke, Ectasy		19	
RandomForest	LSD, Mushrooms, Legalh		52	Gini
	Nicotine, Cannabis		52	Gini
	Meth, VSA, Ketamine, Heroin, Crack, AMyl		111	entropy
	Coke, Ectasy		170	Gini