Compte-rendu Projet SY09 - Drug Consumption

Sébastien Dam, Jinshan Guo, Yassine Oumni

June 2021

Abstract

Cette étude évalue les risques de consommation de différentes drogues. Le jeu de données est intitulé "Drug Consumption" et a été récupéré sur le site UCI Machine Learning Repository. Il présente diverses informations liées à des individus issus de différents milieux sociaux. En particulier, ce travail présente les données par le biais d'une analyse exploratoire avant d'étudier différentes méthodes de classification pour obtenir un modèle susceptible de prédire les risques de consommation des drogues selon la personnalité et les informations démographiques d'un individu. Après des résultats mitigés sur la classification des données brutes, l'utilisation de méthodes portées sur les classes nous permet d'obtenir des précisions et des sensibilités supérieures à 70% et à 85%. Le résultat permet alors de classer en deux catégories les consommateurs ou non-consommateurs pour chaque drogue.

1 Introduction

Ce projet SY09 a pour but d'étudier le jeu de données "Drug Consumption". Ce dernier contient les informations qualitatives et quantitatives de 1885 individus ayant renseigné leurs attributs demographiques ainsi que leurs mesures de personnalité. De plus, les participants ont indiqué la dernière fois où ils ont consommé les 18 drogues légales ou illégales. Le but de ce projet est de créer le meilleur modèle capable d'évaluer le risque pour un individu de consommer chaque drogue. Pour ce faire, nous commencerons par le prétraitement de données et effectuerons une analyse exploratoire de celles-ci. Puis, il s'agira de trouver les attributs les plus utiles pour notre modèle par la séletion de variables et la réduction de dimensions. Finalement, nous appliquerons les algorithmes de classification vus en cours et effectuerons la sélection des modèles pour déterminer le meilleur modèle. Grâce au rééquillibrage des classes, nous arrivons à obtenir des bons résultats de précision et sensibilité. L'obtention d'un modèle efficace nous permettra de prédire le risque de consommation de drogue pour un nouvel individu renseignant les mêmes informations.

2 Description des données

Le jeu de données comporte 1885 enregistrements et 32 colonnes. Pour chaque individu, 12 attributs et 18 drogues sont recensés dont :

1. Cinq colonnes démographiques : l'âge, le genre, le niveau d'éducation, le pays de résidence et l'ethnie ;
2. Sept mesures de personnalité :
   (a) Nscore (N) : névrose, une tendance à long terme à éprouver des émotions négatives telles que la nervosité, la tension, l'anxiété et la dépression ;
   (b) Escore (E) : extroversion, se manifeste par des caractéristiques extraverties, chaleureuses, actives, affirmées, bavardes, joyeuses et à la recherche de stimulation ;
   (c) Oscore (O) : ouverture à l'expérience, une appréciation générale de l'art, des idées inhabituelles et des intérêts imaginatifs, créatifs, non conventionnels et vastes ;
   (d) Ascore (A) : agréabilité, une dimension des relations interpersonnelles, caractérisée par l'altruisme, la confiance, la modestie, la gentillesse, la compassion et la coopération ;
   (e) Cscore (C) : conscienciosité, une tendance à être organisée et fiable, volontaire, persévérant, fiable et efficace ;
   (f) Impulsivité (Imp) mesuré par le BIS-11 ;
   (g) SS : recherche de sensations (en anglais sensation seeking) mesuré par ImpSS.
3. 18 drogues avec la fréquence de consommation indiquée : Alcool, Amphétamines, Nitrite d'amyle, Benzodiazépine, Caféine, Cannabis, Chocolat, Cocaïne, Crack, Ecstase, Héroïne, Kétamine, Drogues légales, Diéthylamide de l'acide lysergique, Méthadone, Champignons magiques, Nicotine, Drogue fictive Semer (Semer est une drogue

fictive qui a été introduite pour identifier les personnes qui sélectionne trop de drogues), Abus de substances volatiles.

Pour chacun des 12 attributs, les données recueillies ont été quantifiées au préalable tandis que chaque drogue a été étiquetée par les valeurs attribuées à leur signification :

1. CL0 : Jamais utilisée
2. CL1 : Utilisée il y a plus d'une décennie
3. CL2 : Utilisée au cours de la dernière décennie
4. CL3 : Utilisée l'année dernière
5. CL4 : Utilisée au cours du mois dernier
6. CL5 : Utilisée la semaine dernière
7. CL6 : Utilisée au cours du dernier jour

La Table 1 recense les statistiques sur les variables qualitatives de notre jeu de données.

TABLE 1 – Répartition des valeurs qualitatives

Variable	Valeurs	Proportion
Age	18-24	34.1%
	25-34	25.5%
	35-44	18.9%
	45-54	15.6%
	55-64	4.9%
	65+	1.0%
Gender	Male	50%
	Female	
Education	Left school before 16	1.5%
	Left school at 16	5.3%
	Left school at 17	1.6%
	Left school at 18	5.3%
	Some college or university	26.8%
	Professional diploma	14.3%
	University degree	25.5%
	Masters degree	15.0%
	Doctorate degree	4.7%
Country	USA	29.5%
	New Zealand	0.3%
	Other	6.2%
	Australia	2.9%
	Republic of Ireland	1.1%
	Canada	4.6%
	UK	55.4%
Ethnicity	Black	1.7%
	Asian	1.4%
	White	91.2%
	Mixed-White/Black	1.1%
	Other	3.3%
	Mixed-White/Asian	1.1%
	Mixed-Black/Asian	0.2%

3 Prétraitement des données

3.1 Inspection des données

D'après la Figure 1, il n'y a pas de valeurs manquantes dans notre jeu de données car toutes les barres dans la figure (générée avec la librairie *missingno*) ci-dessous sont noires, donc pleines.

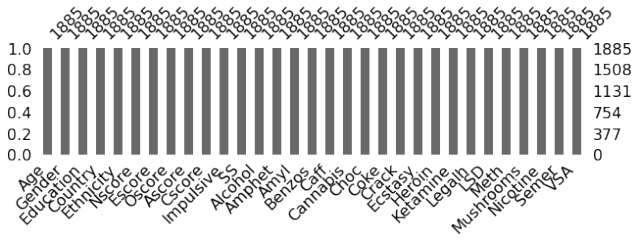

FIGURE 1 – Inspection des valeurs manquantes

3.2 Transformation des données quantitatives

Avant d'établir une analyse exploratoire, il a d'abord fallu associer aux différents attributs quantifiés leur signification disponible sur UCI Machine Learning Repository (par exemple pour l'âge, -0.95197 correspond à la tranche d'âge 18-24 ans, -0.07854 à 24-35 ans...).

3.3 Centrage-réduction des données quantitatives

Comme il y a un grand écart d'échelle parmi les traits de personnalité, il a fallu effectuer le centrage-réduction pour que les variables aient la même importance pour le modèle. Cette transformation est par ailleurs nécessaire pour l'ACP et certains algorithmes.

3.4 Binarisation des classes

En raison du nombre assez important d'étiquettes pour les drogues, nous avons décidé de binariser les classes pour obtenir deux étiquettes en *Non utilisateur* et *Utilisateur* au lieu de sept. Il y a quatre façons de définir ces deux catégories selon les différents seuils de fréquence : basé sur la décennie (Decade-based), basé sur l'année (Year-based), basé sur le mois (Month-based), basé sur la semaine (Week-based). Néanmoins, peu importe la définition choisie, le ratio de *Non utilisateur* et

d'*Utilisateur* pour toutes les drogues est toujours déséquilibré, en particulier pour la définition 'basé sur la semaine'.

Pour faciliter notre analyse, nous avons choisi de binariser les étiquettes en fonction de l'année :

1. *Non utilisateur* regroupe Jamais utilisée (CL0), Utilisée il y a plus d'une décennie (CL1), Utlisée au cours de la dernière décennie (CL2) ;
2. *Utilisateur* regroupe Utilisée l'année dernière (CL3), Utilisée au cours du mois dernier (CL4), Utilisée la semaine dernière (CL5), Utilisée au cours du dernier jour (CL6).

La Table 2 présente la proportion des *Utilisateurs* selon les quatres définitions de binarisation.

	Decade -based	Year -based	Month -based	Week -based
Alcohol	96%	93%	82%	67%
Amphet	36%	23%	13%	9%
Amyl	20%	7%	2%	1%
Benzos	41%	28%	16%	9%
Caff	98%	97%	94%	88%
Cannabis	67%	53%	42%	34%
Choc	98%	98%	95%	79%
Coke	36%	22%	8%	3%
Crack	10%	4%	1%	1%
Ecstasy	40%	27%	13%	4%
Heroin	11%	6%	3%	2%
Ketamine	19%	11%	4%	2%
Legalh	40%	30%	13%	7%
LSD	30%	20%	9%	4%
Meth	22%	17%	9%	6%
Mushrooms	37%	23%	8%	2%
Nicotine	67%	56%	46%	41%
VSA	12%	5%	2%	1%

TABLE 2 – Proportion d'*Utilisateur* après la binarisation en fonction de la décennie, de l'année, du mois et de la semaine

4 Analyse exploratoire

4.1 Analyse monovariée

4.1.1 Aperçu des attributs démographiques

Dans les variables qualitatives Country et Ethnicity, nous avons remarqué un fort déséquilibre dans la répartition des modalités. En effet, USA et UK représentent respectivement 55.4% et 29.5% des pays de résidence tandis que les autres modalités sont à des proportions inférieures à 7%. De même, 91.2% des individus intérrogés ont la couleur de peau blanche, entraînant une proportion très faible des autres ethnies. Par conséquent, nous pensons que le pays de résidence peut avoir un impact dans le risque de consommation de drogue ou non mais le manque de données nous amène à porter une attention particulière à cette variable, tout comme l'ethnie. C'est pourquoi nous effectuerons une sélection de variables dans la section 5.

4.1.2 Aperçu des mesures de personnalité

Nous avons effectué une analyse monovariée des mesures de personnalité en regroupant les valeurs dans des boîtes à moustaches sur la Figure 2 pour Alcohol, Choc, Cannabis et Coke.

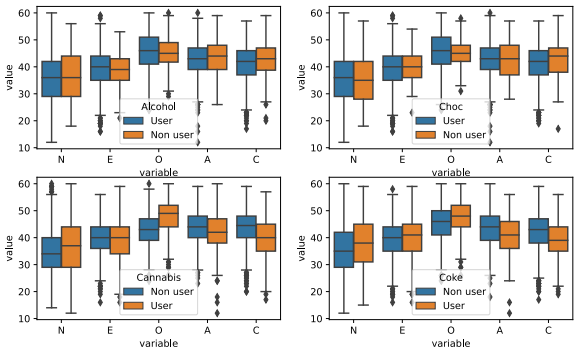

FIGURE 2 – Distribution des mesures de personnalité

Nous constatons que pour la majorité des drogues, hormis Alcohol, Caff et Choc, les utilisateurs possèdent des valeurs médianes de Nscore, Oscore et SS plus élevées que les non-utilisateurs et des valeurs médianes Ascore et Cscore plus faibles.

4.1.3 Aperçu des drogues

La Figure 3 présente la proportion d'utilisateurs de drogues. On observe qu'hormis la Nicotine et le Cannabis, les autres drogues sont soit consommés par plus de 90% des individus sondés, soit par moins de 30%, voire moins de 20%. Nous verrons dans la Section 6 le problème que ce déséquilibre pose et les moyens mis en place pour y pallier.

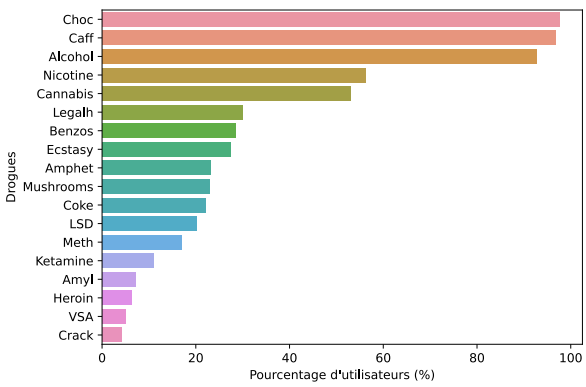

FIGURE 3 – Proportion d'utilisateurs selon les drogues

4.2 Analyse multivariée

4.2.1 Corrélation entre les mesures de personnalité et la consommation de drogues

Pour connaître la corrélation entre les mesures de personnalité et la consommation de drogues, nous avons réalisé un tableau de corrélation non-linéaire de Spearman [1] sur la Figure 4.

FIGURE 4 – Analyse de corrélation de Spearman entre les traits de personnalité et la consommation de drogue

L'alcool, la caféine et le chocolat sont très mal ou pas du tout corrélés aux mesures de personnalité. La fréquence de leur utilisation ne semble donc pas dépendre de la personnalité. On peut donc émettre l'hypothèse 1 suivante : *ne pas considérer ces drogues.*

L'Escore est très mal ou pas du tout corrélé à l'utilisation de toutes les drogues.

Le Ascore est faiblement ou pas du tout corrélé avec toutes les drogues. De plus, nous pouvons observer que la corrélation du Cscore avec les drogues est très similaire à celle d'Ascore, mais est mieux représentée car ses valeurs sont plus élevées. La corrélation la plus forte est avec le Cannabis, le Legalh, l'Amphétamine, l'Ecstasy, la Nicotine.

De même, nous constatons que la corrélation de SS avec les drogues est très similaire à celle de l'Impulsive, mais est mieux représentée car ses valeurs sont plus élevées.

Par conséquent, nous pouvons déduire que les consommateurs des drogues Amphet, Benzos, Cannabis, Coke, Ecstasy, Legalh, LSD, Mushrooms, et Nicotine sont susceptibles d'avoir des Nscore, Oscore et SS plus élevés et un Cscore plus faible. D'une part, cette analyse nous permet de confirmer les constatations de la Figure 2. D'autre part, nous pouvons émettre l'hypothèse 2 selon laquelle *les variables les plus significatives seraient Nscore, Oscore, SS et Cscore.* Nous allons par la suite démontrer cela par les méthodes d'analyse non supervisée avec l'ACP et la CAH.

4.3 Application de l'ACP

Pour vérifier la première hypothèse citée précédemment dans la sous-section 4.2, nous appliquons l'ACP sur les variables de mesures de personnalité et nous gardons les quatre premières composantes principales. Ces dernières expliquent environ 90% de la variance. Les résultats de cette ACP sont résumés dans la Table 3.

TABLE 3 – ACP sur les mesures de personnalité

Composante principale	1	2	3	4
Variance expliquée (%)	43.74	18.82	15.19	12.34

La Figure 5 ci-après présente les résultats obtenus après application de l'ACP sur les drogues Alcohol, Caff, Ecstasy et Mushrooms dans le premier plan factoriel. Dans la représentation des différents plans factoriels pour l'Alcohol, la Caff et le Choc, les points sont très superposés entre eux. Cela montre que les variables initiales n'étaient pas corrélées entre elles dans le cadre de l'étude de ces drogues. En revanche, pour les autres drogues, on observe des points qui sont plus éloignés les uns des autres, ce qui confirme notre hypothèse d'enlever les drogues Alcohol, Caff et Choc. Nous pouvons

interpréter cela comme le fait que ces trois drogues sont très populaires peu importe la personnalité ou les informations démographiques des individus. Par conséquent, l'analyse de ces drogues n'est pas bénéfique pour notre étude.

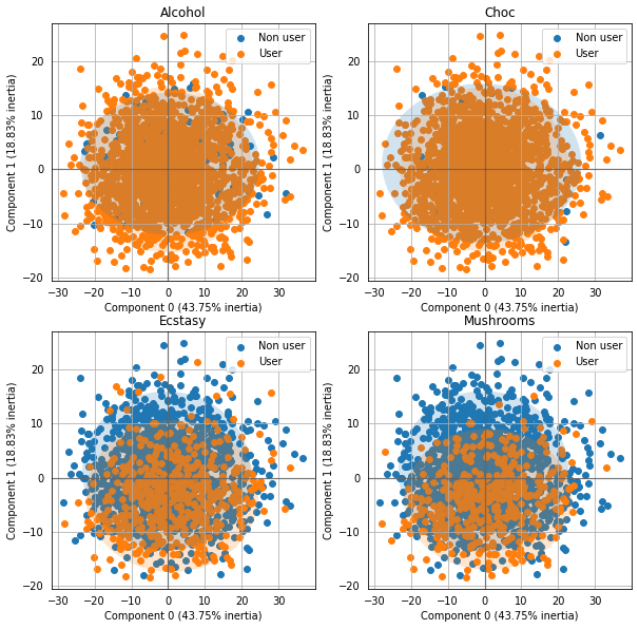

FIGURE 5 – Projection des individus dans le premier plan factoriel pour l'Alcohol, la Caff, l'Ecstasy et les Mushrooms

4.4 CAH

Pour vérifier la deuxième hypothèse citée précédemment dans la sous-section 4.2, nous appliquons la CAH avec le critère de Ward pour observer combien de clusters nous obtenons. Le résultat est présenté dans la Figure 6.

Nous observons quatre clusters, ce qui confirme bien l'utilisation privilégiée des variables énoncées dans l'hypothèse 4.2.1.

5 Sélection de variables

Etant donné que nous avons déjà confirmé l'hypothèse 4.2.1 grâce à l'analyse de corrélation dans la section précédente, cela nous permet de sélectionner les attributs des mesures de personnalité pour l'apprentissage du modèle. Cependant, il nous manque encore les attributs démographiques à choisir. Pour faire ce choix, nous avons créé notre propre méthode décrite dans la

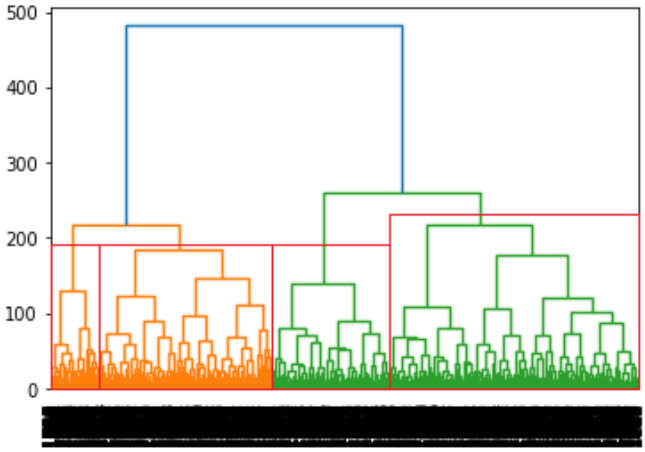

FIGURE 6 – CAH

sous-section 8.2 en prenant en compte la régularisation L1 et L2 en même temps, nous avons ensuite utilisé le Meta-transformer *SelectFromModel* proposé par Scikit-learn pour entraîner un classifieur qui nous permet d'atteinndre cet objectif. Le terme L1 permet de réduire la dimension en supprimant les variables ayant des coefficients nuls, et le terme L2 permet d'équilibrer la compléxité et la capacité de génération du modèle.

La Table 4 présente les attributs démographiques (Age, Gender, Education - Edu, Country - Count, Ethnicity - Eth) à garder (indiqué par le symbole T) pour chaque drogue après application de la méthode de séléction des variables.

	Age	Gender	Edu	Count	Eth
Amphet	T		T	T	
Amyl	T				
Benzos	T		T	T	
Cannabis	T	T	T	T	
Coke	T			T	
Crack				T	
Ecstasy	T	T		T	
Heroin				T	
Ketamine	T				
Legalh	T	T	T	T	
LSD	T			T	
Meth	T		T	T	
Mushrooms	T	T		T	
Nicotine	T	T	T	T	
VSA	T			T	

TABLE 4 – Sélection des attributs démographiques pour chaque drogue. T signifie que que l'on garde l'attribut dans notre modèle.

D'après les résultats ci-dessus, on peut constater qu'Ethnicity n'aura aucun impact quelque soit la drogue et Gender va seulement influencer le Canabis, l'Ecstasy, le Legalh, le Mushrooms et la Nicotine.

6 Evaluation du risque de consommation de drogue

Nous avons utilisé différentes méthodes de classification supervisée pour construire notre modèle. La validation de nos modèles et le choix du meilleur classifieur pour chaque drogue dépend de la forme des classes. En effet, pour les classes où la proportion des labels était fortement déséquilibrée, nous avons utilisé plusieurs mesures pour évaluer la qualité des modèles dont la courbe ROC qui affiche la proportion de vrais positifs en fonction de la proportion de faux positifs. Il s'agit donc de l'affichage de la proportion de prédictions correctes pour la classe positive (le long de l'axe des ordonnées) en fonction de la proportion de prédictions incorrectes de la classe négative (le long de l'axe des abscisses). Nous avons également utilisé la sensibilité définie par le ratio $\frac{tp}{tp+fn}$, où tp est le nombre de vrais positifs et fn le nombre de faux négatifs. Cette mesure permet de donner la probabilité qu'une personne soit sujette à consommer une drogue sachant qu'elle est réellement utilisatrice de la drogue. Ainsi, la sensibilité nous paraît important pour détecter correctement les éventuels consommateurs de drogues. Pour les classes où la proportion d'étiquettes était équilibrée, nous avons utilisé la méthode de k-fold cross-validation pour obtenir la précision moyenne entre les différents plis. Au départ, la plupart des classes était déséquilibrée car la majorité des individus intérrogés ne consommaient pas de drogues (hormis Alcohol, Caff, Choc). Nous expliquerons dans cette section nos démarches pour pallier à ce problème.

6.1 Réduction des dimensions

Au vu du grand nombre de variables qualitatives et quantitatives, nous avons décidé d'effectuer une ACP pour tenter de réduire celui-ci.

Etant donné que les données étaient déjà quantifiées, elles pouvaient être considérées comme des valeurs réelles. Nous avons donc considéré ces valeurs sans prétraitement dans un premier temps. Nous avons obtenu des résultats peu satisfaisants car il a fallu dix composantes principales pour expliquer 97,75% de la variance. Cela n'était pas très fructueux car les variables initiales étaient au nombre de 12.

Dans un deuxième temps, nous avons donc décidé de reprendre les données avec leur signification de base (hormis pour Impulsivity et SS) pour appliquer la quantification des données manuellement. Pour cela, nous avons effectué une transformation des variables qualitatives en variables binaires en appliquant un *codage disjonctif complet* sur les variables nominales et un *codage additif* sur la variable ordinale Age. Il en résultait 33 variables. Ainsi, après application de l'ACP sur ces dernières, nous avons obtenu un résultat bien meilleur car il a suffi de cinq composantes principales seulement pour expliquer 98.24% de la variance. Les résultats de cette ACP sont résumés dans la Table 5.

Table 5 – ACP sur les données quantifiées manuellement

Composante principale	1	2	3	4
Variance expliquée (%)	43.16	18.63	15.01	12.17

6.2 Résultats de la classification sur les classes

En utilisant les classifieurs Régression Logistique (RL), K-plus proches voisins (K-NN), arbre de décision (DT), support vector machines (SVM), Naive (NB), Analyse Discriminante Linéaire (ADL), Analyse Discriminante Quadratique (ADQ) et Forêt Aléatoire (RF), nous obtenons les résultats affichés dans la Table 6.2. Seul le meilleur classifieur est retenu pour chaque drogue et la qualité des classifieur est mesuré par la précision : $\frac{vp}{vp+fp}$ où vp est le nombre de vrais positifs et fp le nombre de faux positifs. Nous avons choisi ce ratio car il montre la capacité du modèle à correctement prédire les vrais positifs parmi toutes les prédictions positives réalisées. C'est une mesure efficace lorsque les classes sont déséquilibrées, ce qui est notre cas.

Nous avons délibéremment gardé les drogues Alcohol, Caff et Choc dans ces résultats pour montrer leur score élevé (trop). Cela est dû au fait que les classes sont fortement déséquilibrées. En effet, les modèles de classification essayent de maximiser la précision en prédisant arbitrairement le fait que les individus ne consomment pas de drogues. De ce fait, les classifieurs que nous obtenons n'ont quasiment aucun pouvoir de prédiction. En revanche, les drogues pour lesquels nous obtenons des précisions acceptables entre 70 et 80% sont le Cannabis, la Nicotine et le Legalh. D'après la Figure 3, il s'agit des drogues qui ont une proportion d'utilisateurs et non-utilisateurs assez équilibrée (35 à 60% d'utilisateurs). En revanche, pour les autres drogues, la précision est inférieure ou égale à 60%. Pour pallier à cela, nous avons

Table 6 – Résultats des classifications sur les drogues

Drogue	Meilleur classifieur	Précision (%)
Alcohol	QDA	93.19
Amphet	LDA	54.74
Amyl	DT	13.33
Benzos	K-NN	59.26
Caff	NB	96.80
Cannabis	RL	83.90
Choc	DT	98.18
Coke	LDA	51.28
Crack	RL	9.85
Ecstasy	RF	60.00
Heroin	DT	16.00
Ketamine	RL	26.36
Legalh	RF	79.41
LSD	LDA	52.59
Meth	K-NN	60.00
Mushrooms	LDA	56.35
Nicotine	RL	74.93
VSA	RL	14.35

utilisé deux méthodes de correction de classes déséquilibrées : le resampling [3] et le "clustering" des drogues.

6.3 Resampling

Au vu du nombre d'individus que nous avons dans le jeu de données, nous avons choisi la méthode SMOTE pour augmenter le nombre d'observations dans les drogues où la part des non-utilisateurs est faible. L'idée de base de l'algorithme SMOTE est d'analyser et de simuler quelques types d'échantillons, et d'ajouter de nouveaux échantillons simulés artificiellement à l'ensemble de données, de sorte à ce que les catégories des données d'origine ne soient plus sévèrement déséquilibrées. Cette méthode est basée sur la méthode des K-NN pour construire les nouvelles observations. En effet, l'algorithme va prendre un point au hasard de la classe des non-utilisateurs. Puis, parmi les k plus proches points, un de ces derniers va être sélectionné et un nouveau point va être créé entre ces deux pour constituer le nouvel individu. Les résultats de la classification après resampling sont présentés dans la Figure 6.3. Le meilleur classifieur est choisi tout d'abord selon la sensibilité la plus grande et en suite en fonction de la précision la plus grande.

Nous constatons bien que le K-NN marche toujours mieux que les autres classifieurs pour toutes les drogues sauf Cannabis, Ecstasy et Nicotine, qui sont mieux classifées par ADL, RF et SVM rescpectivement. Les trois premières composantes principales présentent au moins

Table 7 – Résultats de classification sur les drogues après resampling

Drogue	Meilleur Classifieur	Précison (%)	Sensibilité (%)
Amphet	K-NN	76.95	91.56
Amyl	K-NN	80.62	92.81
Benzos	K-NN	70.57	85.19
Cannabis	LDA	80.19	78.96
Coke	K-NN	74.07	88.24
Crack	K-NN	87.23	97.05
Ecstasy	RF	70.14	88.42
Heroin	K-NN	85.23	97.88
Ketamine	K-NN	81.75	95.72
Legalh	K-NN	77.35	88.76
LSD	K-NN	80.04	91.16
Meth	K-NN	78.62	92.11
Mushrooms	K-NN	80.15	94.78
Nicotine	SVM	66.14	73.53
VSA	K-NN	85.78	95.64

plus de 90% d'inertie expliquée totale. Les données sont linéairement séparables et en faible dimension, c'est pourquoi le K-NN est un bon choix dans le plupart de cas.

6.4 "Clustering" des drogues

Nous appelons "clustering" des drogues le regroupement des drogues en différents clusters. Les clusters sont créés selon la corrélation entre les drogues. Chaque cluster contient un centre : une drogue qui possède les meilleures corrélations avec les autres drogues de son cluster.

Pour créer les clusters, nous avons commencé par étudier la corrélation entre les drogues.

6.4.1 Corrélation entre les drogues et choix des clusters

Pour ce faire, nous avons utilisé la corrélation non-linéaire de Spearman car les modalités de consommation de drogues sont ordonnées (Utilisée il y a plus d'une décennie $<$ Utilisée au cours de la dernière décennie $<$ Utilisée l'année dernière $<$ ...). Nous obtenons les corrélations sur la Figure 7.

Nous choisissons les clusters de manière à ce que le nombre de drogues contenus dans un cluster soit le plus faible possible dans un autre cluster. Cela revient à choisir des clusters les plus différents possibles. Par exemple, Amphet possède de bonnes corrélations non-linéaires avec les autres drogues donc il ne serait pas judicieux de

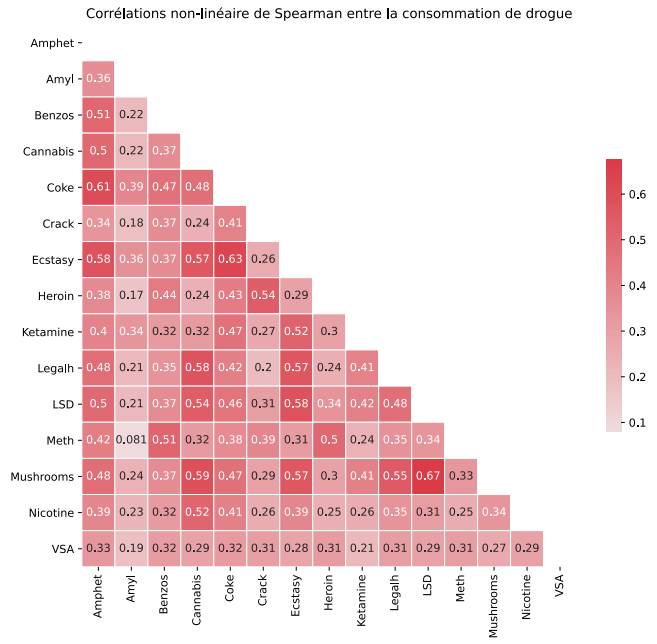

FIGURE 7 – Corrélations non-linéaire de Spearman pour la consommation de drogues

le choisir comme cluster car il est susceptible de contenir une grande part de drogues identiques aux autres potentiels clusters. D'après la Figure 7, nous pouvons relever trois drogues qui possèdent des corrélations non-linéaires assez élevées avec les autres drogues : *Benzos*, *Ecstasy* et *Heroin*. Nous considérons donc les trois clusters suivants :

1. BenzosCl : Amphet, Meth, Coke et Benzos
2. EcstasyCl : Amphet, Cannabis, Coke, Mushrooms, LSD, Legalh, Ketamine et Ecstasy
3. HeroinCl : Coke, Crack, Meth et Heroin

En utilisant ces trois clusters au lieu des drogues initiales, nous avons vérifié que les tendances pour les traits de personnalités Nscore, Oscore, Cscore et SS étaient similaires à celles que nous avions trouvé dans la sous-section 4.1, et c'est bien le cas.

Ce clustering a l'avantage d'équilibrer les nouveaux clusters de drogues obtenus. Nous résumons la part d'utilisateurs et de non-utilisateurs de drogues dans la Table 8. Nous observons que les utilisateurs sont beaucoup mieux répartis, hormis pour HeroinCl.

Etant donné que HeroinCl ne possède que 31.03% d'utilisateurs, nous considérons que la classe est déséquilibrée et effectuons alors en plus un resampling sur celle-ci pour notre classification.

TABLE 8 – Part d'utilisateurs des clusters

Cluster	BenzoCl	EcstasyCl	HeroinCl
Utilisateur (%)	44.03	57.77	31.03

6.5 Résultats de la classification sur les clusters

La sélection de variables citée dans la section 5 utilisée sur les clusters de drogues donnent le même résultat pour chacun d'entre eux : tous les attributs démographiques sont importants hormis Ethnicity.

En utilisant les mêmes classifieurs que dans la sous-section 6.2, nous obtenons les précisions et sensibilités résumées dans la Table 9.

TABLE 9 – Résultats des classifications sur les clusters de drogues

Cluster	Meilleur Classifieur	Précision (%)	Sensibilité (%)
BenzoCl	RL	51.34	78.46
EcstasyCl	ADL	83.83	82.28
HeroinCl	RL	63.84	72.76

Nous obtenons de biens meilleurs résultats pour la précision. Ce résultat était attendu grâce au rééquilibrage des classes.

7 Conclusion

Dans cette étude, nous avons analysé les variables susceptibles d'influencer le risque de consommation de drogues d'un individu. A partir du jeu de données, les facteurs les plus importants semblent être les mesures de personalité, en particulier la névrose, l'ouverture à l'expérience, la conscienciosité et la recherche de sensations, ainsi que les informations démographiques tels que l'âge, le genre et l'éducation. Grâce aux méthodes non supervisées comme l'ACP ou la CAH, nous avons pu vérifier nos hypothèses et avons selectionné les variables et les drogues les plus importantes dans notre étude, au sens qu'elles nous donnent les meilleures informations. Nous obtenons des classifieurs rendant des scores de précisions de plus de 70%, ce qui est plutôt satisfaisant. Pour obtenir ces résultats, nous avons procédé au rééquilibrage des consommations de drogues en regroupant les drogues corrélées non-linéairement et en utilisant la méthode d'undersampling.

Pour avoir des meilleurs résultats, plusieurs pistes sont à envisager. Nous avons effectué une binarisation des classes pour obtenir de meilleurs résultats avec nos modèles. Il existe d'autres méthodes de classification dans le Deep Learning pour effectuer la prédiction sur des problèmes multi-classes, mais nous n'avons pas encore les connaissances pour cela.

8 Annexes

8.1 Fonction pour visualiser les corrélations

```
def heatmap_corr(df, method='spearman', already=False, mask=True, nominal=False, title=None, figsize=(20,9)):
    ''' Fonction pour visualiser la correlation
    df - peut etre des donn es pures (sans correlations calculees) ou un DataFrame avec des correlations d j calculees (dans ce cas, l'attribut already doit etre defini sur True)
    method - 'Pearson'(lineaire) ou 'Spearman'(non-lineaire)
    mask - pour cacher les valeurs en triangulaire sup rieur et manquantes
    nominal - pour les corr lations de donnees nominales, les valeurs sont dans la plage (0, 1) au lieu de (-1, -1) et nominal = True doit tre suivi par already = True
    '''
    if not already:
        corr = df.corr(method=method)
    elif already:
        corr = df
    cmap = sns.diverging_palette(220, 10, as_cmap=True)
    vmax = corr.max().max()
    if nominal:
        center = 0.5
        cmap=None
    elif not nominal:
        center = 0
    if mask:
       mask = np.zeros_like(corr, dtype=np.bool)
       mask[np.triu_indices_from(mask)] = True
       vmax = corr.replace(1, -2).max().max()
    elif not mask:
      mask=None
    f, ax = plt.subplots(figsize=figsize)
    sns.heatmap(corr, cmap=cmap, mask=mask, vmax=vmax, center=center, annot=True, linewidths=.5, cbar_kws={'shrink': 0.5})
    if title:
      plt.title(title)
    plt.show()
```

8.2 Fonction pour sélectionner les attributs en combinant la régularisation L1 et L2

```
from sklearn.linear_model import LogisticRegression
class LR(LogisticRegression):
  def __init__(self, threshold=0.01, dual=False, tol=1e-4, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='liblinear', max_iter=100, multi_class='ovr', verbose=0, warm_start=False, n_jobs=1):
    self.threshold = threshold
    LogisticRegression.__init__(self, penalty='l1', dual=dual, tol=tol, C=C, fit_intercept=fit_intercept, intercept_scaling=intercept_scaling, class_weight=class_weight, random_state=random_state, solver=solver, max_iter=max_iter, multi_class=multi_class, verbose=verbose, warm_start=warm_start, n_jobs=n_jobs)
    self.l2 = LogisticRegression(penalty='l2', dual=dual, tol=tol, C=C, fit_intercept=fit_intercept, intercept_scaling=intercept_scaling, class_weight = class_weight, random_state=random_state, solver=solver, max_iter=max_iter, multi_class=multi_class, verbose=verbose, warm_start=warm_start, n_jobs=n_jobs)

  def fit(self, X, y, sample_weight=None):
    super(LR, self).fit(X, y, sample_weight=sample_weight)
    self.coef_old_ = self.coef_.copy()
    self.l2.fit(X, y, sample_weight=sample_weight)

    selected_column_list = []

    cntOfRow, cntOfCol = self.coef_.shape
    for i in range(cntOfRow):
      for j in range(cntOfCol):
        coef = self.coef_[i][j]
        if coef != 0:
          idx = [j]
          coef1 = self.l2.coef_[i][j]
          for k in range(cntOfCol):
            coef2 = self.l2.coef_[i][k]
            if abs(coef1-coef2) < self.threshold and j != k and self.coef_[i][k] == 0:
              idx.append(k)
          mean = coef / len(idx)
          self.coef_[i][idx] = mean

    for i, item in enumerate(self.coef_[0]):
      if item != 0:
        selected_column_list.append(X.columns[i])

    print("Importante features selected: ", selected_column_list)

    return self
```

Références

[1] Juhi Ramzai. Clearly explained : Pearson V/S Spearman Correlation Coefficient. *Disponible en ligne à cette URL.*

[2] Jaiganesh Nagidi. Best Ways To Handle Imbalanced Data In Machine Learning. *Disponible en ligne à cette URL.*

[3] Nitesh V. Chawla. SMOTE : Synthetic Minority Over-sampling Technique. *Disponible en ligne à cette URL.*