MRHIZOU ABDESSAMAD (2121028)
KASSIR AYOUB (2060605)
OUTALEB EL HOUSSINE (2105335)

École Polytechnique de Montréal
Département de Mathématiques et Génie Industriel
IND6212 Exploration de données industrielles

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

Analyse de l'influence du tabac, de l'alcool et vapotage sur la consommation du cannabis

Contexte et problématique

De multiples études et analyses permettent d'encadrer les politiques publiques qui veillent sur la consommation de diverses substances, telles que la consommation d'alcool, le cannabis, le tabagisme et l'usage du vapotage. En 2021, Statistique Canada a mené une enquête sur les tendances de consommation du tabac et de la nicotine, selon différents groupes d'âge dans l'ensemble du territoire Canadien.

À travers cette enquête, nous cherchons les éventuels liens entre ces tendances de consommation. Depuis la légalisation du cannabis au Canada, sa consommation est de plus en plus repandue chez des personnes de différents groupes d'âge, particulièrement chez les jeunes. Donc, dans ce projet d'étude, on s'intéresse à la consommation du cannabis à l'aide d'un problème de classification, pour **prédire si une personne est susceptible de consommer le cannabis dans sa vie, ainsi les facteurs qui influencent le plus sur cette consommation**, afin de fournir une meilleure compréhension qui pourrait aider à réduire les risques liés à la dépendance créée par celui-ci.

La base des données

La base de données que nous avons utilisée dans le cadre de ce projet, regroupe les informations sur les tendances de consommation des différentes substances pour des personnes regroupées, par 5 catégories d'âge (25-34 ; 35-44 ; 45-54 ; 55-64 ; 65 et plus), genre, taille du ménage (nombre de personnes dans chaque logement) et les habitudes de consommation. La base de données contient 99 variables (colonnes) et 8113 observations (lignes). Puisqu'il s'agit d'une enquête, l'ECTN[1] a créé des variables du type catégorique, qui permettent à l'usager de répondre directement à l'enquête par des réponses directes préparées au préalable. L'ensemble du jeu de données ne contiendra donc pas de valeurs aberrantes puisqu'on n'étudie que des variables catégoriques de plages prédéfinies.

Tableau 1Extrait de la BD

Variable	Description	Type de données	Format
PUMFID	Numéro d'ordre généré de façon aléatoire pour le fichier principal	Discret	Numérique
HHLDSIZE	Taille du ménage	Discret	Numérique
GENDER	Genre du répondant	Discret	Numérique
TBC_05AR	A fumé cigarettes - vie	Discret	Numérique
TBC_05BR	Âge a fumé la première fois - cigarette complète	Discret	Numérique
TBC_10AR	Fréquence a fumé cigarettes - 30 jours	Discret	Numérique
TBC_10BR	Nb de jours a fumé cigarettes au moins 1 fois par semaine - 30 jours	Discret	Numérique
TBC_15R	A fumé au moins 100 cigarettes - vie	Discret	Numérique
TBC_20R	Quand cesser de fumer cigarette	Discret	Numérique

Préparation de données

1. Réduction des données

L'étude consiste à classifier si une personne est susceptible de consommer ou non de cannabis. Le seul critère, lié aux tendances des consommations de cannabis, qui est pertinent dans notre cas, est la variable CAN_05AR (A fumé cannabis – vie). Cette variable nous a servis comme une base d'entrainement de notre modèle. Les autres variables liées au cannabis, tel que celles qui traitent les

[1] La BD est extraite du STATISTIC CANADA (voir lien dans le paragraphe Références), la source fournit la base des données ainsi qu'un document qui détaille et décrit les différentes critères (variables) de cette enquête.

fréquences de consommation, la provenance du cannabis … n'auront aucune utilité dans notre étude. Donc, ces variables seront retirées du dataset. Pour cela, on a filtré sous Python toutes les variables qui comportent le mot « cannabis » sauf la variable CAN_05AR qui représente la réponse des observations de notre modèle.
Nous avons supprimé aussi les variables qui n'ont aucun rapport avec la prédiction de la consommation du cannabis chez un individu, telles que : l'ID de l'observation ('PUMFID'), la date de collecte de chaque observation ('VERDATE'), et le poids d'enquête de la personne ('WTPP').
La variable province de résidence ('PROV_C') semble d'être un critère significatif, or, l'enquête a été menée sur l'ensemble du territoire canadien, avec des poids des provinces qui ne sont pas bien répartis (39,6% Ontario ; 22,8% Québec ; 3% Manitoba …). Donc on a opté à éliminer cette variable de notre étude.

2. Nettoyage des données

Dans les 8113 observations de l'enquête, 3144 personnes ont déjà consommé du cannabis (code 1), contre 4956 qui ne l'ont jamais fait (code 2). Or, 12 personnes n'ont pas répondu à cette question (code 9). Et puisque nous avons besoin de faire une classification, et étant donné que cela représente moins de 0,014% nous avons décidé de supprimer ces 12 observations associées. Ensuite, pour simplifier le codage, nous avons changé les codes de cette variable de (1 ;2) à (0 ;1).

3. Données aberrantes

Notre jeu de données ne contient pas des valeurs aberrantes dans les colonnes des variables qui peuvent influencer notre variable de réponse (consommation du cannabis). Les différentes valeurs présentes sont comprises dans des intervalles déjà connus.

4. Codage des variables

Nous avons remarqué que le jeu de données sur lequel nous travaillons contient des variables qui sont codés 6 (96 pour d'autres variables) pour certaines observations avec une mention **valid_skip**. Ces observations sont codées de cette manière puisqu'elles ont déjà été traitées au préalable par le fournisseur du jeu de données. La mention **valid_skip** veut dire que la question (variable) a été sautée pour l'observation en question puisqu'elle ne s'applique pas à la situation du répondant, par exemple pour une personne qui a répondu « non » à la variable « A fumé cigarettes – vie », elle ne pourra pas répondre à des questions comme la fréquence de consommation du tabac etc…
Et vu, que le nombre d'observations qui contiennent ces mentions pour différentes variables représente dans certains cas jusqu'à la moitié des observations, nous nous trouvons dans l'obligent de garder cette mention pour les raisons suivantes :

- La fiabilité des poids de chaque variable vis-à-vis de l'ensemble des observations.
- Ne pas avoir le souci de la manière du traitement de telle situation pour les éventuelles prédictions.
- Garantir que les algorithmes feront l'apprentissage correctement.

Tandis que si on les supprime notre jeu de données ne contiendront que très peu de données et donc on aura des modèles biaisés.

5. Etude de la corrélation par la mesure de CRAMER'S V :

Après avoir éliminé les différentes variables qu'on a jugées non pertinentes pour notre classification, nous avons eu encore 88 variables. Parmi ces variables, il y a celle qui ne sont pas bien corrélés par rapport aux autres, c'est pour cela, nous avons étudié la corrélation entre ces variables. Et puisque

nos variables sont des facteurs catégoriques, plus formellement, nous avons besoin d'une méthode qui permette de mesurer les associations entre chaque deux facteurs catégoriques. Nous avons utilisé **la méthode Cramer V**, qui est basée sur une variation nominale du test **Khi Deux**. Comme pour la corrélation, la sortie est dans l'intervalle de [0,1], où 0 signifie aucune association et 1 est une association complète. (Contrairement à la corrélation, il n'y a pas de valeurs négatives, car il n'y a pas d'association négative. Soit il y en a, soit il n'y en a pas).

Tableau 2Matrice de corrélation des variables

	HHLDSIZE	GENDER	TBC_05AR	TBC_05BR	TBC_10AR	TBC_10BR	TBC_15R	TBC_20R	TBC_25QR	TBC_30AR	...	CAN_30FR	CAN_30GR
HHLDSIZE	1.000000	0.040855	0.198911	0.128803	0.147717	0.029248	0.168329	0.109507	0.000000	0.061897	...	0.101182	0.102591
GENDER	0.040855	1.000000	0.050427	0.047069	0.055100	0.025499	0.051298	0.052898	0.100299	0.044875	...	0.070693	0.066874
TBC_05AR	0.198911	0.050427	1.000000	0.741494	0.999876	0.716030	0.825388	0.684785	0.432902	0.365057	...	0.116779	0.121996
TBC_05BR	0.128803	0.047069	0.741494	1.000000	0.526559	0.158421	0.603026	0.295929	0.101071	0.145193	...	0.100437	0.102165
TBC_10AR	0.147717	0.055100	0.999876	0.526559	1.000000	0.695805	0.693081	0.529599	0.308230	0.570936	...	0.124809	0.127833
...	...	...	...	...	...	...	...	...	...	...	...	...	...
ALC_10	0.057208	0.083663	0.166199	0.091766	0.137971	0.061785	0.160319	0.058704	0.026247	0.084572	...	0.174481	0.172833
AGEGROUP	0.302365	0.049315	0.267148	0.164272	0.198915	0.038696	0.233959	0.178312	0.037056	0.059037	...	0.156229	0.158649
DV_SSR	0.157502	0.048752	0.694263	0.474425	0.796389	0.432945	0.812853	0.633788	0.233905	0.567077	...	0.069588	0.075217
DV_VP30R	0.176578	0.061963	0.097838	0.106987	0.145942	0.123484	0.094710	0.151099	0.073982	0.147351	...	0.217975	0.219412
DV_ALC30	0.086579	0.011290	0.148528	0.150216	0.156154	0.049278	0.166323	0.084206	0.035913	0.066702	...	0.113308	0.113427

82 rows × 82 columns

Cette méthode nous a donné une matrice de corrélation des 82 variables. Chaque paire de variables avec une valeur supérieure à 0,95 signifie que ces deux variables sont quasi-semblables, donc nous nous contentons d'une des deux variables.
Finalement, nous avons gardé **78 variables**.
La base de données finale avec laquelle nous avons entrainé notre modèle est fournie avec le rapport.

Classification et évaluation des modèles

Après avoir préparé les données, nous passons maintenant à l'étape de la conception du modèle de classification. À cet effet, la base de données a été divisée en deux ensembles : entraînement et test (avec **test_size=0.33**). Ceci nous permettra d'éviter le surapprentissage lors de l'entrainement des modèles. Nous utiliserons **la métrique de l'accuracy** pour évaluer la performance des modèles qui peut se calculer automatiquement sur python.

Comme nous nous trouvons devant un problème de classification, différents modèles et méthodes peuvent êtres envisageables. Afin d'identifier le modèle qui servira au mieux pour notre étude, nous avons utilisé les modèles suivants : arbre de décision, réseaux de neurones et réseaux bayésiens, Méthode des k plus proches voisins et les SVM. Et nous avons obtenu les scores ci-dessous :

MODÈLE	Decision Tree	Random Forest	Neural Net	Nearest Neighbors	Naive B ayes	Linear SVM	RBF SVM
ACCURACY	0.7460	0.7807	0.7853	0.785260	0.7501	0.7834	0.7475

En regard des différents résultats obtenus, si nous devions orienter notre choix sur un modèle prédictif plutôt qu'un autre, nous opterions davantage pour **les forêts aléatoires** puisque la performance de ce modèle à prédire de la consommation du cannabis fait partie des meilleurs résultats obtenus. Ainsi, avec un choix optimal des différentes hyperparamètres, on obtiendra des performances plus hautes que celles obtenues. D'autres part, le modèle de forêts aléatoires nous permet de voir aussi, quelles sont les variables qui influencent la variable de sortie. Et donc, va nous servir à identifier les facteurs qui impactent le plus la consommation du cannabis. D'une autre part, les arbres de décision sont les

méthodes les plus faciles à implémenter, or, pour un seul arbre de décision, la performance reste fortement dépendante de l'échantillon de données de départ. Pour cette raison, il est plus avantageux d'opter pour un grand nombre d'arbre afin d'obtenir un modèle plus robuste et d'éviter les surajustements. Donc, pour le reste de notre étude, nous avons analysé nos données avec la méthode des **forêts aléatoires.**

1. Modèle forêts aléatoires : ParameterGrid

Pour obtenir de meilleurs résultats de l'algorithme forêts aléatoires, il est très important de bien choisir les hyperparamètres du modèle. Ces hyperparamètres peuvent être ajustés pour optimiser les performances. Dans le cas d'une forêt aléatoire, les hyperparamètres incluent le nombre d'arbres de décision dans la forêt et le nombre de caractéristiques considérées par chaque arbre lors de la division d'un nœud. Afin d'identifier les hyperparamètres optimaux, nous avons appliqué le "Random Hyperparameter Grid" sur l'ensemble d'entrainement, avec le paramétrage initial ci-dessous :
*param_grid = { **'n_estimators'**: [200, 500, 1000], **'max_features'**: ['sqrt', 'log2'], **'max_depth'** : [5, 10, 25], **'min_samples_leaf'**: [1, 2, 4], **'min_samples_split'**: [2, 5, 10], **'criterion'** :['gini']}*
Pour chaque itération, l'algorithme a calculé la précision (accuracy) pour chaque combinaison possible (3x2x3x3x3x1=162 combinaisons), et nous avons obtenu le meilleur score : **0,79823** avec les paramètres optimaux suivants :
{**'criterion'**: 'gini', **'max_depth'**: 10, **'max_features'**: 'sqrt', **'min_samples_leaf'**: 1, **'min_samples_split'**: 2, **'n_estimators'**: 500})

2. Modèle forêts aléatoires optimal

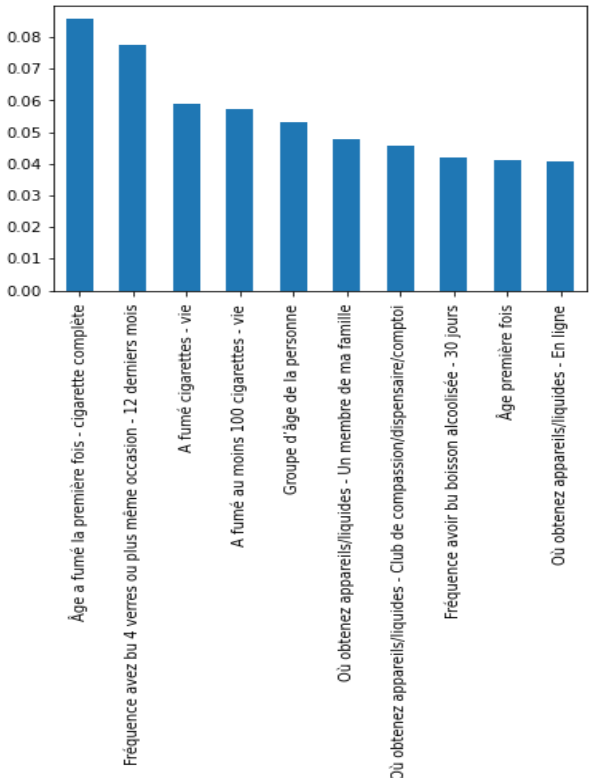
Nous avons appliqué ensuite le modèle des forêts aléatoires avec les hyperparamètres optimaux sur l'ensemble d'entrainement, après, on a calculé le score du modèle sur l'ensemble du test. Le score de notre modèle est **0.79723.** Ce score indique que notre modèle a réussi de correctement prédire à l'environ de 80% des observations tests. Cela peut, dans un premier lieu, supporter notre hypothèse sur l'influence des tendances de consommation des autres substances sur la consommation du cannabis.

3. Importance des variables selon le modèle des forêts aléatoires

Dans cette étape, nous avons essayé d'étudier l'importance des variables de notre base de données selon le modèle des forêts aléatoires qu'on a conçu.
A partie du diagramme à barre, on constate que les variables : L'âge dans lequel la personne a fumé sa première cigarette (TBC_05BR) et la variable : Fréquence de consommation excessive de l'alcool (ALC_10), constituent les variables les plus significatives selon notre modèle pour prédire si la personne est susceptible de consommer du cannabis.

Ce résultat apparaât tout à fait logique si on fait la comparaison avec ce qu'on peut constater dans la vie quotidienne. Une grande partie des personnes qui ont commencé à fumé dès un jeune âge, elles ont consommé au moins une fois les cannabis dans leur vie. La même chose peut être dite sur l'effet de l'alcool. Une consommation excessive pourrait mener aux tentations de tester du cannabis.
On constate aussi le facteur de l'âge de la personne, qui impacte sur la subtilité de la consommation du cannabis.
Or, les facteurs liés au vapotage sont peu présents. Une chose qui peut apparaître aussi normale, vu que la plupart des gens qui vapotent, l'utilisent comme un substitut des autres substances et pas l'inverse.

Figure 1 Diagramme à barres

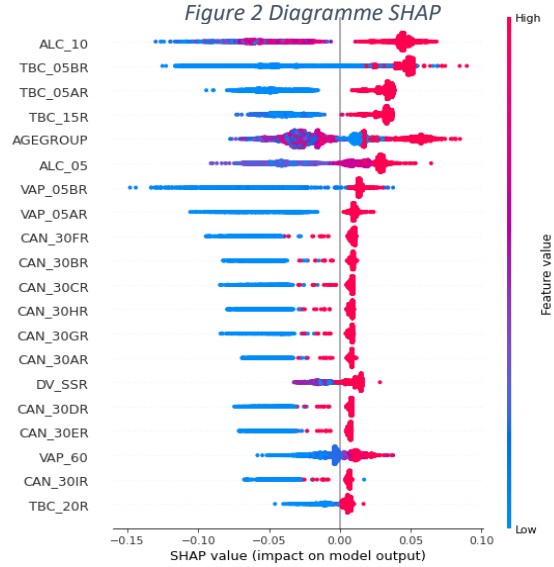


4. Shap tree

L'utilisation du SHAP_Value dans l'explication du modèle, nous a permis de mesurer la contribution des caractéristiques d'entrée aux prédictions individuelles. Parmi ses avantages, les valeurs SHAP montrent non seulement l'importance de la fonctionnalité, mais également si la fonctionnalité a un impact positif ou négatif sur les prédictions.

Les résultats du modèle SHAP montrent que la plupart des valeurs SHAP élevées (points rouges) sont des valeurs positives pour les différentes variables, par exemple pour la variable « ALC_30 » les valeurs du SHAP élevées sont positives, cela signifie que la personne qui consomme excessivement de l'alcool a plus de chance d'avoir fumé du cannabis. Ce résultat est convenable avec le résultat qu'on a obtenu à travers l'algorithme des forêts aléatoires.

Figure 2 Diagramme SHAP



Conclusion

Depuis la légalisation de cette substance au Canada, la consommation du cannabis est de plus en plus fréquente parmi les gens de différents groupes d'âge. Notre objectif à travers cette étude, est de voir est ce que cette consommation est influencée par d'autres consommations du tabac, le vapotage et l'alcool.

Nous avons commencé notre analyse par faire une préparation de données pour qu'elles soient prêtes à l'utilisation par la suite.

Après on est passé à l'étape de l'implémentation des modèles, pour pouvoir comparer la performance de différents modèles, on a dressé un tableau qui permet de voir le score donné par chaque méthode. Nous avons trouvé que les différents modèles ont des scores très proches avec un léger écart en faveur des **forêts aléatoires** (score de **0.79723** pour les données test).

En plus d'être le modèle qui donne le meilleur résultat, le modèle de forêts aléatoires nous permet de voir aussi, quelles sont les variables qui influencent le plus la consommation du cannabis. Il a l'avantage, aussi, d'être plus robuste qu'un simple arbre de décision par exemple.

Comme notre objectif est de savoir quelles sont les effets des substances tels que l'alcool, le tabac ou bien l'usage du vapotage, sur la tendance de la consommation du cannabis. Le modèle de forêts aléatoires nous a permis de conclure que **le tabac influence le plus** la consommation du cannabis **suivi par l'alcool** et après par **le vapotage en degré moins faible**.

Voici quelques pistes de réflexions pour approfondir l'étude sur la consommation du cannabis. Peut-on trouver d'autres facteurs qui influencent cette consommation l'entourage de la personne, son éducation et son état psychique ou sa santé mentale etc.

Références

Enquête canadienne sur le tabac et la nicotine (2021-07-12): fichier de microdonnées à grande diffusion:
https://ouvert.canada.ca/data/fr/dataset/7147c035-4418-499c-9623-a0584debfcb1?fbclid=IwAR1ODiqmU3RAtGSxd8aXcJfHg0anrCIy2oPbyIIwqteXzX0Z6j7pP0stdUA

The Search for Categorical Correlation
https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9