Présentation Projet Final

ADJALLA Mylena – SINEYOKO Assitan

Contexte du sujet

Les risques d'addictions aux drogues sont aujourd'hui de plus en plus présents en France et dans le monde en général.

Niveaux de consommation en 2019

L'Observatoire français des drogues et toxicomanies (OFDT) publie régulièrement des statistiques sur les consommations de substances psychoactives. Les données présentées ici ont été publiées en 2019.

Estimation du nombre de consommateurs de substances psychoactives en France métropolitaine parmi les 11-75 ans

	ALCOOL	TABAC	CANNABIS	COCAÏNE	ECSTASY	HÉROÏNE
Expérimentation	47 M	36 M	18 M	2,1 M	1,9 M	500 000
Usage dans l'année	43 M	15 M	5 M	600 000	400 000	-
Usage régulier	9 M	-	1,5 M	-	-	-
Usage quotidien	5 M	13 M	900 000	-	-	-

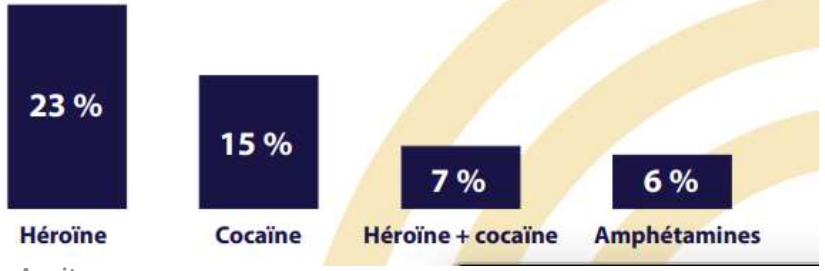
Source : OFDT. Drogues, chiffres clés – 8ème édition. 2019. Téléchargeable sur www.ofdt.fr

- Expérimentation : consommation au moins une fois au cours de leur vie
- Usage dans l'année : consommation au moins une fois dans l'année écoulée
- Usage régulier :
  Alcool : au moins 3 consommations d'alcool dans la semaine
  Cannabis : au moins 10 fois dans le mois

503 décès liés à l'usage abusif de substances illicites ou de médicaments en 2019 (39,1 ans d'âge moyen) (données DRAMES 2019, CEIP-A de Grenoble et ANSM).

Les opioïdes seuls ou en association sont impliqués dans **77 %** de ces décès (données DRAMES 2019, CEIP-A de Grenoble et ANSM).

Principales substances illicites impliquées dans ces 503 décès [29]



Présentation du dataset choisi

Lien du dataset : [UCI Machine Learning Repository: Drug consumption (quantified) Data Set]

Ce dataset présente un ensemble d'informations relatives à la consommation de certaines drogues; Il se compose de 32 attributs et de 1885 individus.

	ID	Age	Sexe	Education_level	Country	Ethnicity	Neuroticism	Extraversionness	Openness	Agreability	Conscientiousness	Impulsiveness	Ser
0	1	0.50	0.48	-0.06	0.96	0.13	0.31	-0.58	-0.58	-0.92	-0.01	-0.22	
1	2	-0.08	-0.48	1.98	0.96	-0.32	-0.68	1.94	1.44	0.76	-0.14	-0.71	
2	3	0.50	-0.48	-0.06	0.96	-0.32	-0.47	0.81	-0.85	-1.62	-1.01	-1.38	
3	4	-0.95	0.48	1.16	0.96	-0.32	-0.15	-0.81	-0.02	0.59	0.58	-1.38	
4	5	0.50	0.48	1.98	0.96	-0.32	0.74	-1.63	-0.45	-0.30	1.31	-0.22	
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1880	1884	-0.95	0.48	-0.61	-0.57	-0.32	-1.19	1.74	1.89	0.76	-1.14	0.88	
1881	1885	-0.95	-0.48	-0.61	-0.57	-0.32	-0.25	1.74	0.58	0.76	-1.52	0.88	
1882	1886	-0.08	0.48	0.45	-0.57	-0.32	1.13	-1.38	-1.28	-1.77	-1.39	0.53	
1883	1887	-0.95	0.48	-0.61	-0.57	-0.32	0.91	-1.92	0.29	-1.62	-2.57	1.29	
1884	1888	-0.95	-0.48	-0.61	0.21	-0.32	-0.47	2.13	1.66	1.11	0.42	0.88	

1885 rows × 32 columns

ısness	Impulsiveness	Sensation_Seeking	Alcohol	Amphet	Amyl	Benzos	Caff	Cannabis	Choc	Coke	Crack	Ecstasy	Heroine	Ketamin	LegalH	LSD
-0.01	-0.22	-1.18	CL5	CL2	CL0	CL2	CL6	CL0	CL5	CL0	CL0	CL0	CL0	CL0	CL0	CL0
-0.14	-0.71	-0.22	CL5	CL2	CL2	CL0	CL6	CL4	CL6	CL3	CL0	CL4	CL0	CL2	CL0	CL2
-1.01	-1.38	0.40	CL6	CL0	CL0	CL0	CL6	CL3	CL4	CL0	CL0	CL0	CL0	CL0	CL0	CL0
0.58	-1.38	-1.18	CL4	CL0	CL0	CL3	CL5	CL2	CL4	CL2	CL0	CL0	CL0	CL2	CL0	CL0
1.31	-0.22	-0.22	CL4	CL1	CL1	CL0	CL6	CL3	CL6	CL0	CL0	CL1	CL0	CL0	CL1	CL0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
-1.14	0.88	1.92	CL5	CL0	CL0	CL0	CL4	CL5	CL4	CL0	CL0	CL0	CL0	CL0	CL3	CL3
-1.52	0.88	0.77	CL5	CL0	CL0	CL0	CL5	CL3	CL4	CL0	CL0	CL2	CL0	CL0	CL3	CL5
-1.39	0.53	-0.53	CL4	CL6	CL5	CL5	CL6	CL6	CL6	CL4	CL0	CL4	CL0	CL2	CL0	CL2
-2.57	1.29	1.22	CL5	CL0	CL0	CL0	CL6	CL6	CL5	CL0	CL0	CL3	CL0	CL0	CL3	CL3
0.42	0.88	1.22	CL4	CL3	CL0	CL3	CL6	CL3	CL6	CL3	CL0	CL3	CL0	CL0	CL3	CL3

Appropriation des données et Visualisations

- Changement des noms de colonnes et mapping pour mettre les valeurs correspondantes
- Mise en avant de la description des données

```
Pays={
-0.09765:"Australie" ,
0.24923:"Canada" ,
-0.46841:"Nouvelle-Zelande" ,
-0.28519:"Autres" ,
0.21128:"République d'Irlande",
0.96082:"Royaume-Uni" ,
-0.57009:"États-Unis"
}
df["Country"]= list(map(lambda x: Pays[x] if x in Pays.keys() else "Vide",df["Country"]))
```

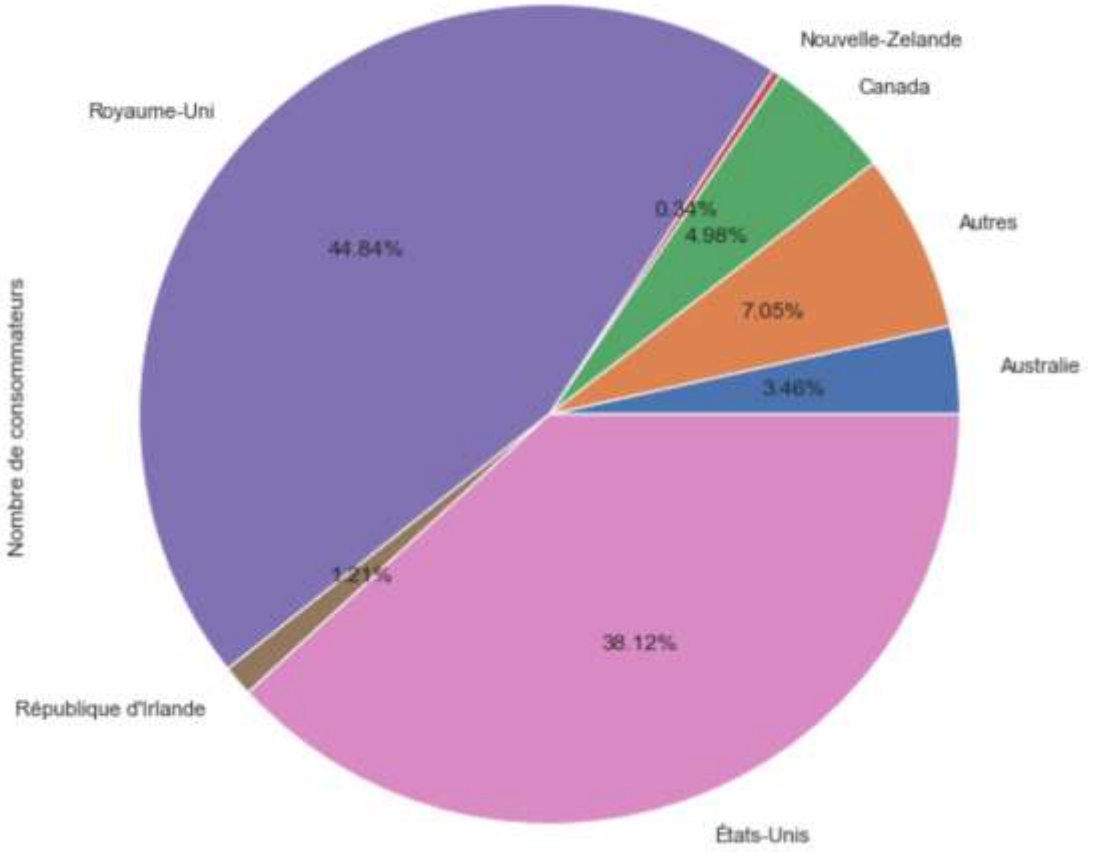
Age	Sexe	Education_level	Country	Ethnicity
0.50	0.48	-0.06	0.96	0.13
-0.08	-0.48	1.98	0.96	-0.32
0.50	-0.48	-0.06	0.96	-0.32
-0.95	0.48	1.16	0.96	-0.32
0.50	0.48	1.98	0.96	-0.32

Age	Sexe	Education_level	Country	Ethnicity
35-44	Femme	Certificat/diplôme professionnel	Royaume-Uni	Blanche/Asiatique
25-34	Masculin	Doctorat	Royaume-Uni	Blanc
35-44	Masculin	Certificat/diplôme professionnel	Royaume-Uni	Blanc
18-24	Femme	Maîtrise	Royaume-Uni	Blanc
35-44	Femme	Doctorat	Royaume-Uni	Blanc

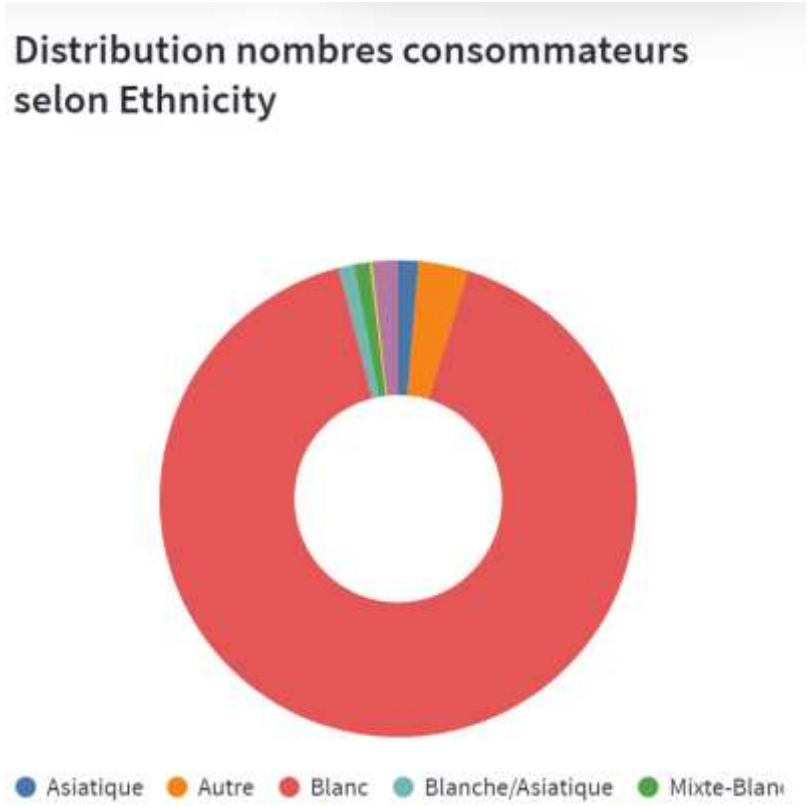
Data visualisation

Visualisation descriptive

- Pourcentage de consommateurs par pays



Data visualisation



Data Visualisation

Visualisation descriptive : Observations

- Le nombre de consommateurs varie selon l'âge et que les individus âgés de 18 à 34 ans représentent la majorité de consommateurs;
- La majorité des consommateurs se trouve aux Etats-Unis et au Royaume Uni (82,96% des consommateurs);
- Les personnes d'ethnie blanche sont les plus consommatrices de drogues dans les différents pays mentionnés dans le dataset;
- Le sexe a en effet une influence sur la consommation de drogues;
- Les différents traits de personnalité, l'âge et le niveau d'éducation influent sur les tendances de consommation;

Objectifs et Problématique

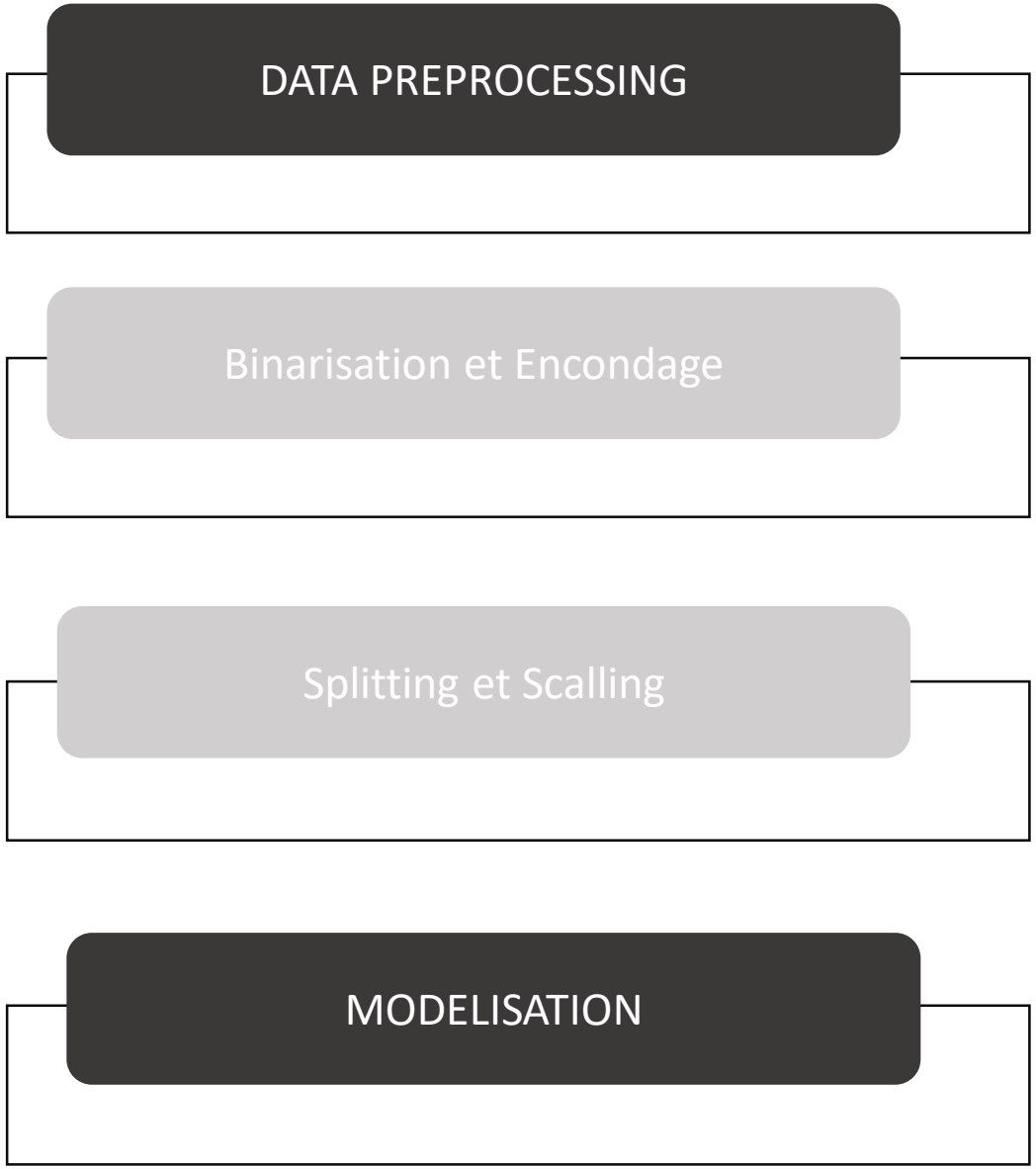
OBJECTIFS

- Participer à la prévention et à la sensibilisation des populations face aux risques d'addictions aux drogues dangereuses;
- Utiliser nos compétences et expériences afin d'accompagner et de soutenir les institutions médicales dans leur lutte contre les addictions

PROBLEMATIQUE

Comment assurer la prédiction du risque de consommation de drogues d'un individu selon les variables explicatives présentes ?

Machine learning



Data Preprocessing

- Encodage

Cannabis	Coke
0	0
1	1
1	0
1	1
1	0
...	...
1	0
1	0
1	1
1	0
1	1

Age	Sexe	Education_level	Country	Ethnicity
2	0	4	4	3
1	1	6	4	2
2	1	4	4	2
0	0	7	4	2
2	0	6	4	2
...	...	...	...	...
0	0	8	6	2
0	1	8	6	2
1	0	5	6	2
0	0	8	6	2
0	1	8	5	2

- Binarisation

Education_level0	Education_level1	Education_level2	Education_level3	Education_level4	Education_level5	Education_level6	Education_level7	Education_level8
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	1	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	1	0	0
...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	1
0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	1

Data Preprocessing

- Création de nouveaux datasets

[Connaître les drogues et leurs effets | Gouvernement du Québec (quebec.ca)]

Selon différentes études toutes les autres drogues en dehors du chocolat et de la caféine sont reconnues comme étant des drogues dangereuses

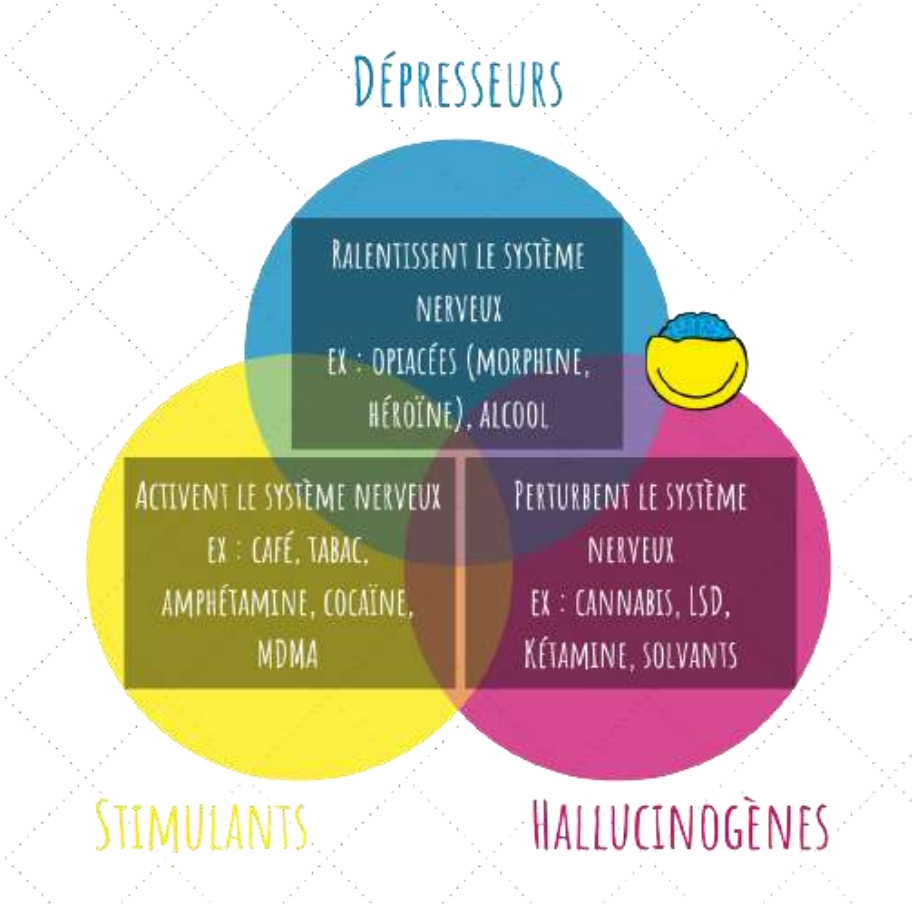
Creation des 3 classes de drogues

```
df_ml["Hallucinogènes"]=df_ml.apply(lambda x: int((x["LSD"]+ x["Mushrooms"]+ x["Ketamin"]+ x["Cannabis"]+ x["LegalH"]+
```

```
df_ml["Stimulants"]=df_ml.apply(lambda x: int((x["Amphet"]+x["Coke"]+x["Ketamin"]+x["LegalH"]+x["Nicotine"]+x["Ecstasy"
```

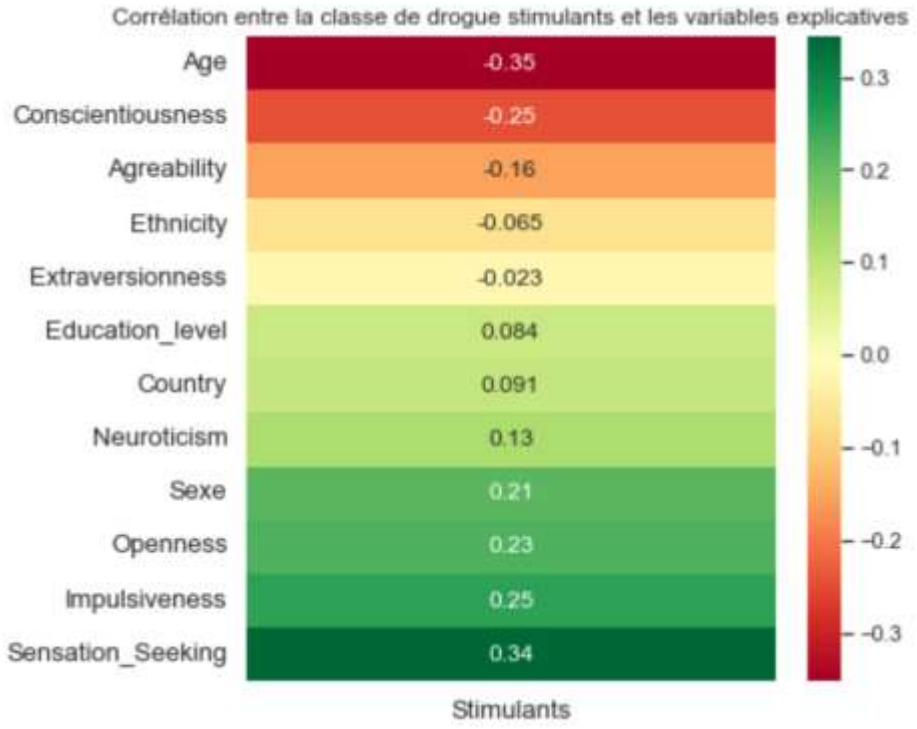
```
df_ml["Depresseurs"]=df_ml.apply(lambda x: int((x["Alcohol"]+x["Heroine"]+x["Meth"]+x["Amyl"]+x["LegalH"]+x["Benzos"]+x
```

Notes: si la valeur vaut 1: User et si 0 : Non user



Observations

- Matrice de corrélation



- On constate que, pour chacune des classes, il existe de très faibles corrélations avec les variables explicatives.(on n'atteint pas les 80% en valeur absolue) ;

- Elimination des variables country(Etats-Unis et Royaume uni et Ehthnicity(Population blanche) en raison de leur surreprésentativité et de faibles valeurs de leur coefficient de corrélation.

Data Preprocessing

• SPLITTING

Train : 80% , Test : 20%

```
# For the Hallucinogen dataset

Hl_X=df_ml_HL.drop("Hallucinogènes",axis=1)
Hl_Y=np.array(df_ml_HL[["Hallucinogènes"]]).reshape(1885,)

# For the Stimulant dataset

St_X=df_ml_ST.drop("Stimulants",axis=1)
St_Y=np.array(df_ml_ST[["Stimulants"]]).reshape(1885,)

# For the Depresseur dataset

Dp_X=df_ml_DP.drop("Depresseurs",axis=1)
Dp_Y=np.array(df_ml_DP[["Depresseurs"]]).reshape(1885,)
```

```
import random

Hl_X_train,Hl_X_test,Hl_Y_train,Hl_Y_test=train_test_split(Hl_X,Hl_Y, test_size=0.2,random_state=100)

St_X_train,St_X_test,St_Y_train,St_Y_test=train_test_split(St_X,St_Y, test_size=0.2,random_state=100)

Dp_X_train,Dp_X_test,Dp_Y_train,Dp_Y_test=train_test_split(Dp_X,Dp_Y, test_size=0.2,random_state=100)
```

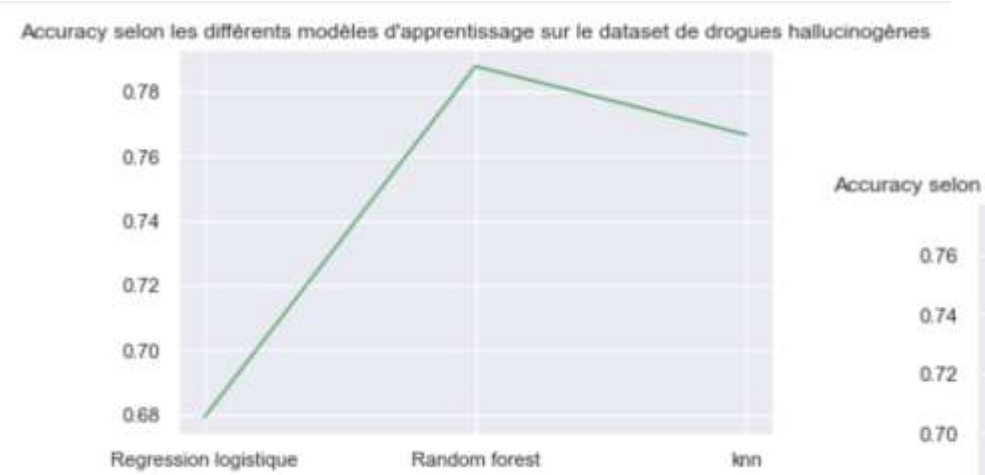
• SCALLING

Afin d'avoir une échelle définit et fixe en min-max pour toutes ces colonnes, nous décidons donc faire un minmax scaler sur ces colonnes.

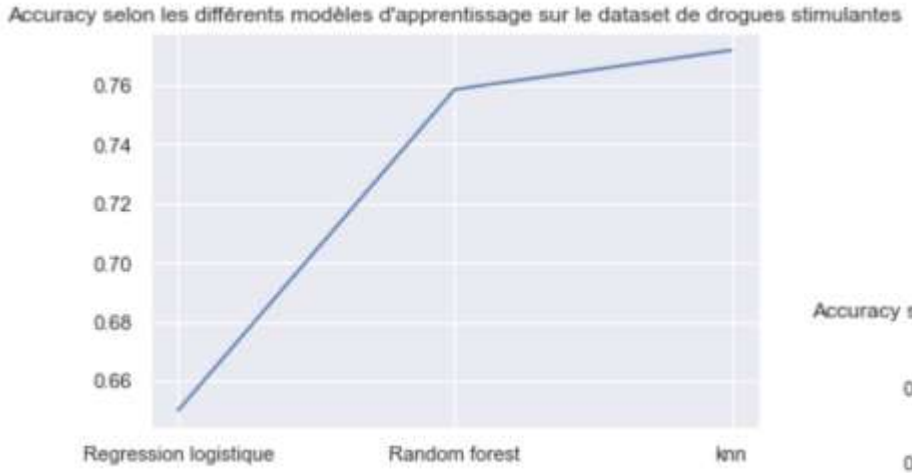
Neuroticism	Extraversionness	Openness	Agreability	Conscientiousness	Impulsiveness	Sensation_Seeking
1508.000000	1508.000000	1508.000000	1508.000000	1508.000000	1508.000000	1508.000000
0.526282	0.500042	0.530768	0.479020	0.475684	0.469903	0.518518
0.167855	0.151520	0.160000	0.150149	0.150463	0.172800	0.241304
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.414613	0.393845	0.413998	0.385251	0.378274	0.337920	0.388117
0.535165	0.500507	0.527023	0.474207	0.475814	0.428474	0.539559
0.633354	0.597404	0.647268	0.591737	0.565147	0.565343	0.710933
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Modélisation

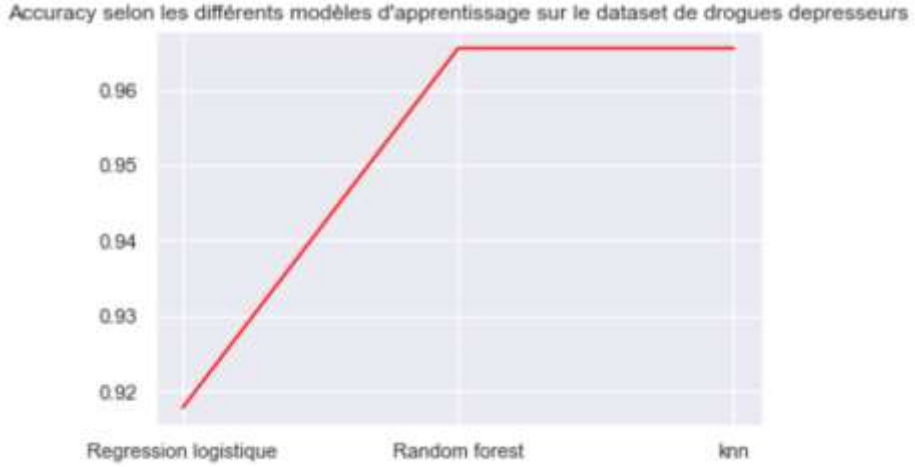
Régression logistique :



Random Forest :



K Nearest Neighbors



Modélisation

CONCLUSION:

- La régression logistique performe moins sur les datasets que les modèles de random forest et de KNN;

- Les scores de prédiction des modèles sont importants sur le dataset des drogues depresseurs.

Cela pourrait sembler être bien, mais relève en fait d'une surreprésentativité de la classe USER (1) dans le target de ce dataset et donc d'un effet d'overfitting.

Nous avons sélectionné le modèle de prédiction Random Forest

Résultat KNN et RF

```
La matrice de confusion est :
 [[  0  13]
 [ 18 346]]
```

Résultat Régression Logistique

```
La matrice de confusion est :
 [[  0  13]
 [  0 364]]
```

API Flask

NOTRE 1ère SOLUTION :

Utiliser les modèles construits pour les 3 groupes de drogues afin de développer un outil de prédiction des risques de consommation, qui pourrait être utilisé par le corps médical.

LES ETAPES DE DEVELOPPEMENT

- Création des modèles avec les phases de preprocessing : PIPELINE
- Exportation des modèles en format pickle;
- Création de l'application d'interface de programmation avec FLASK

EXEMPLE :

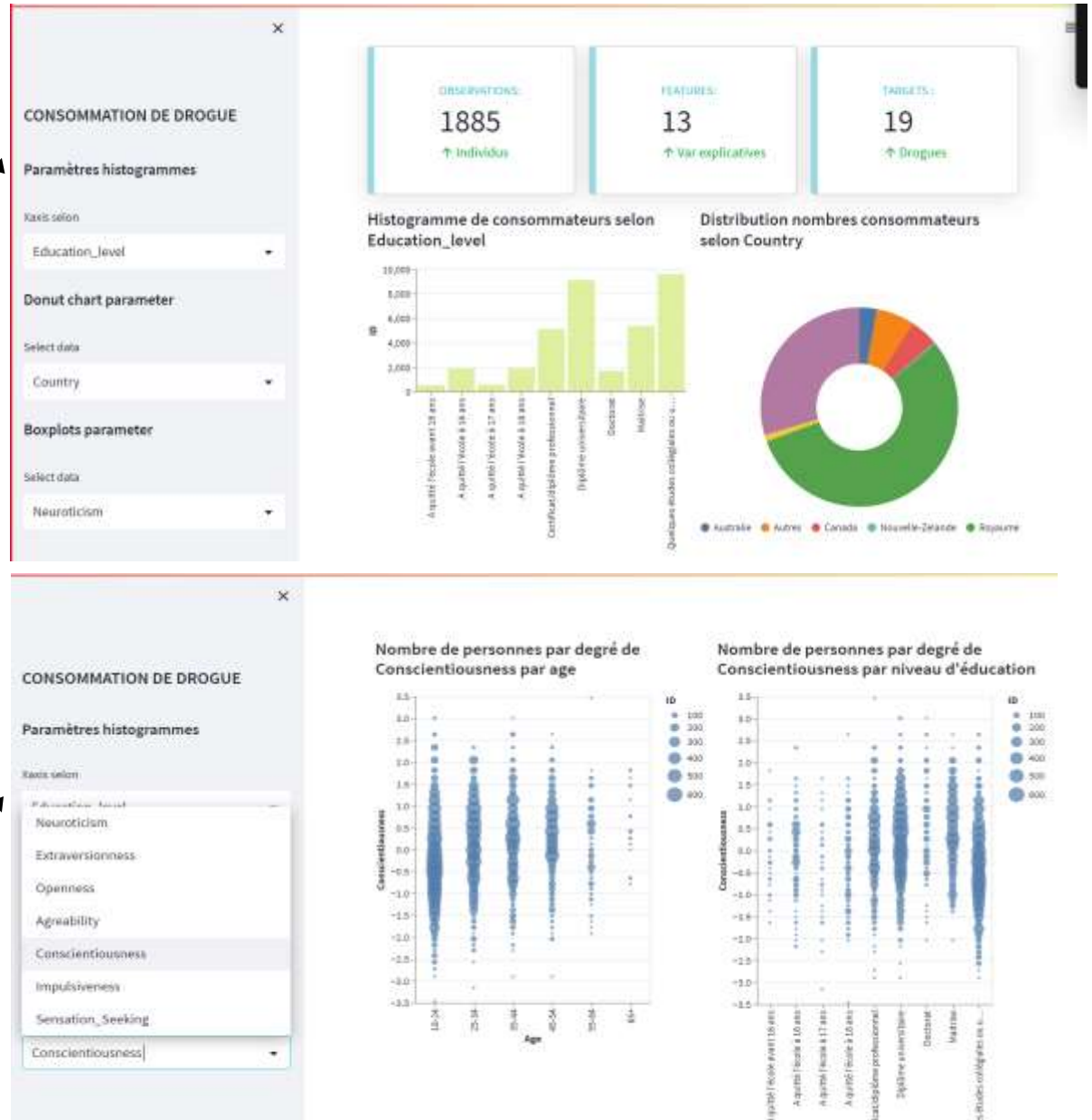
DASHBOARD Streamlit

NOTRE 2ème SOLUTION :

Mettre en place un tableau de bord mettant en avant la distribution du nombre de consommateurs de drogues selon différents facteurs.

L'utilité de ce dernier étant de participer à la sensibilisation des populations sur la consommation de drogues .

Choix de sélection des facteurs



Merci pour votre attention