Drug Consumption

Thomas Osorio-Thanujan Puvikaran

DIA6

Dataset Composition

Information

University study

Medical area

Anonymized data

Patient drugs consumed in his life

Personality study in order to show the effect of drugs consumption

General Composition

Shape: 1880 rows x 30 columns

Features: 11 float, 19 categorical

No Target

No NaN

The Drugs

19 drugs

- including a fictional (Semer) and medicinal (Legal Highs) one

6 levels of use for each drugs

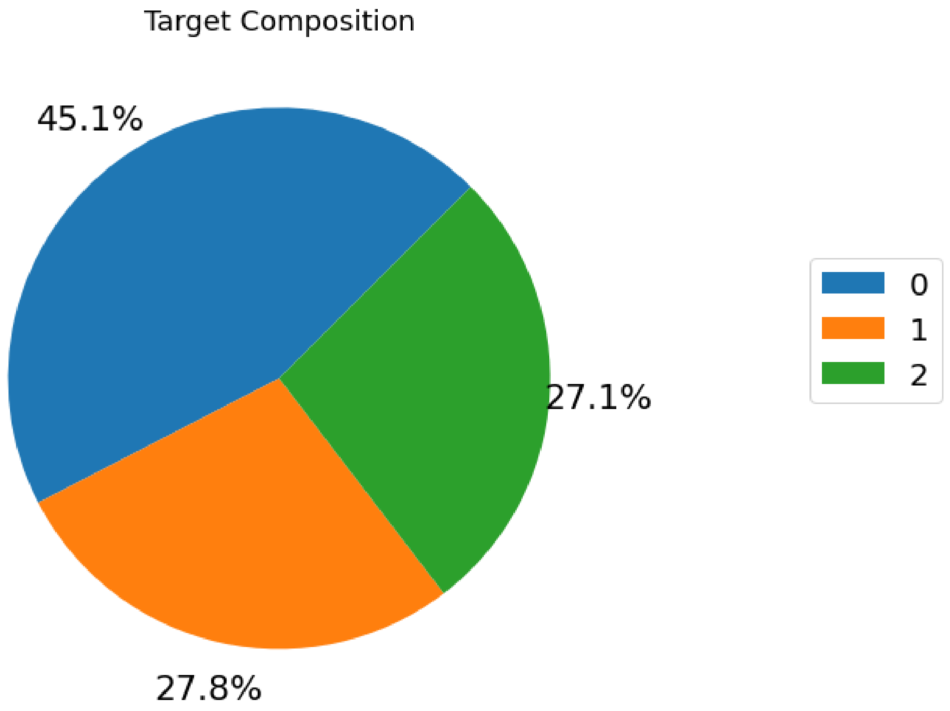
- type: str

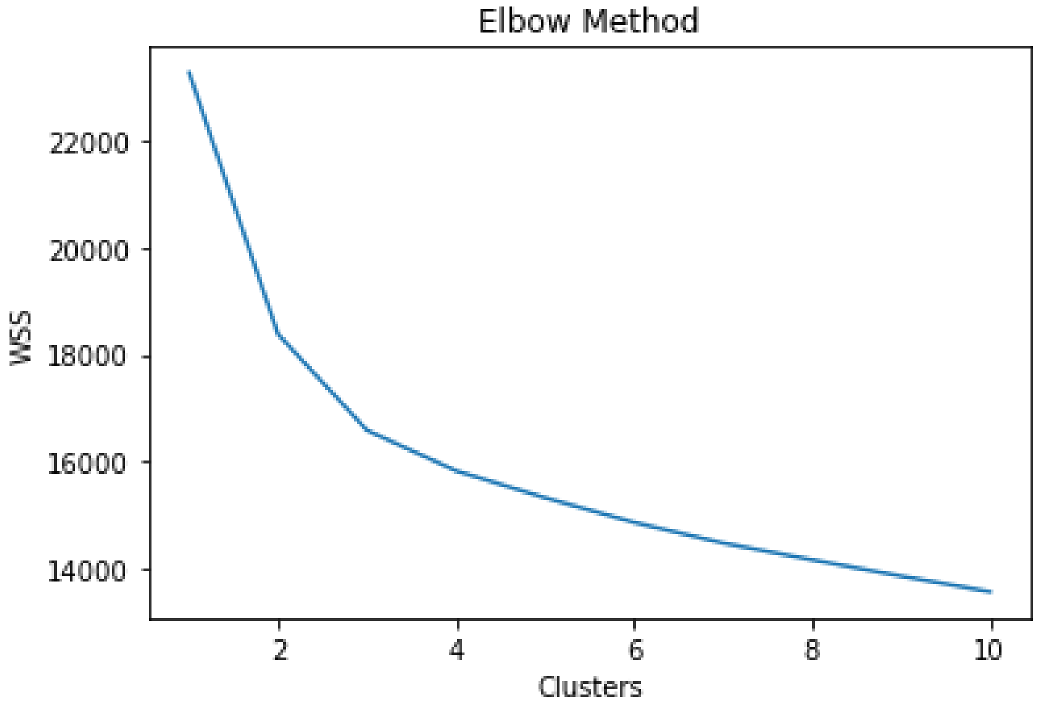
According to the database provider, we can modificate the drug use:

- « CL0 » and « CL1 » as nonuser: we affect 0
- « CL2 » to « CL5 » as user: we affect 1

Target Clustering

- Goal: Define a new target
- Kmeans clustering with Elbow method
  - 3 optimal Cluster
- Unbalanced target clustering
- Now: Discover what are these groups

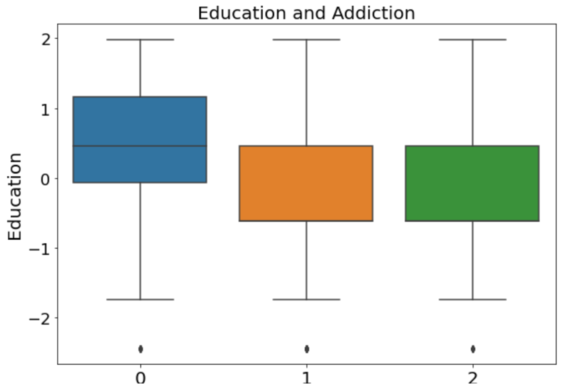


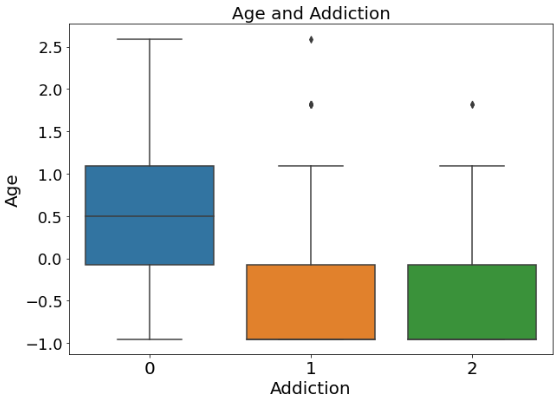


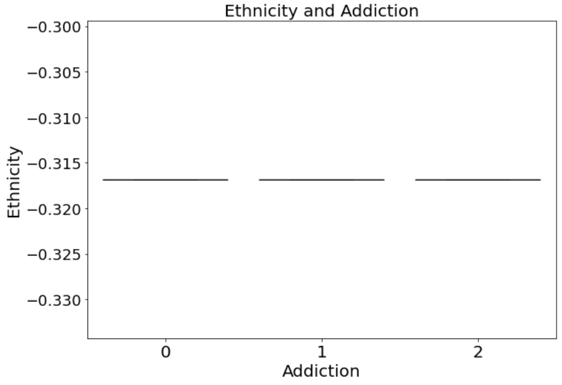
Univariate Analysis with Target

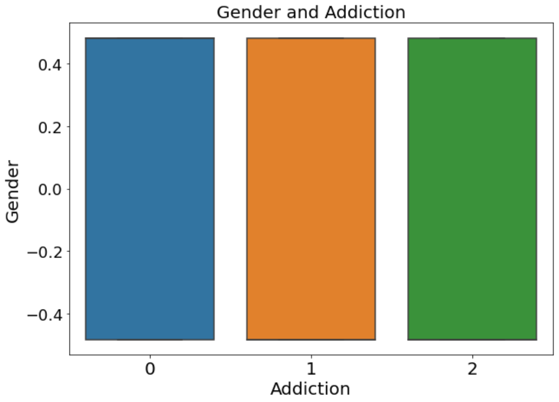
General Features

- All the features are centered and scaled
- Age and Education seems to have an impact
- Ethnicity and Gender don't seem useful



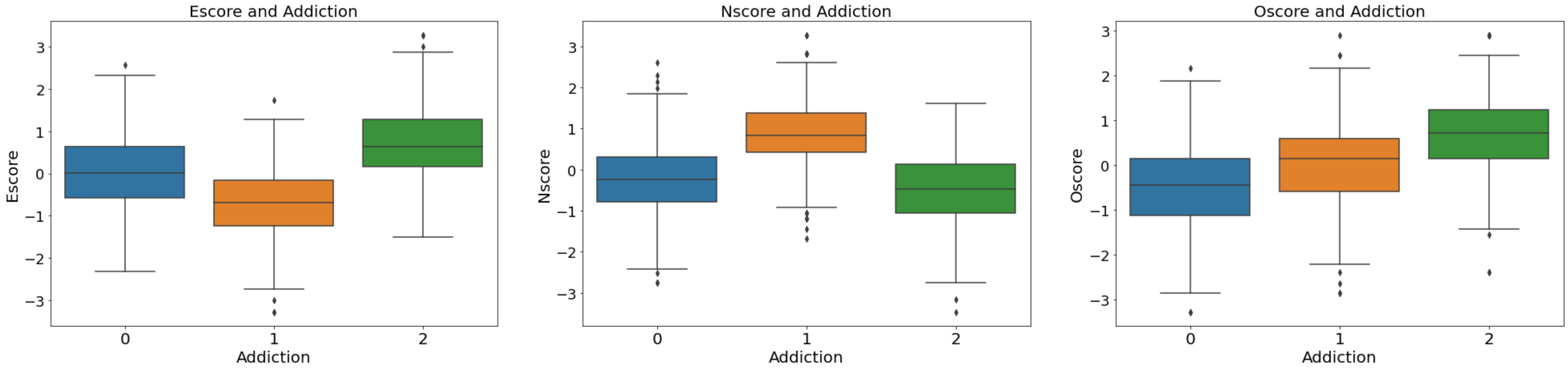






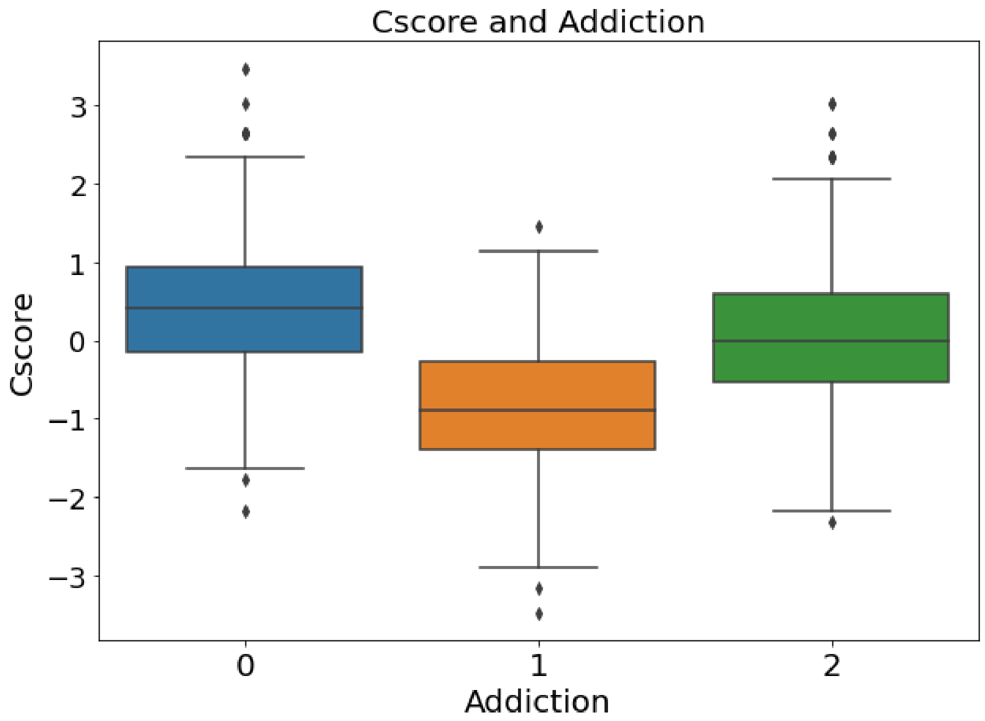
NEO PI-R: Personality Survey

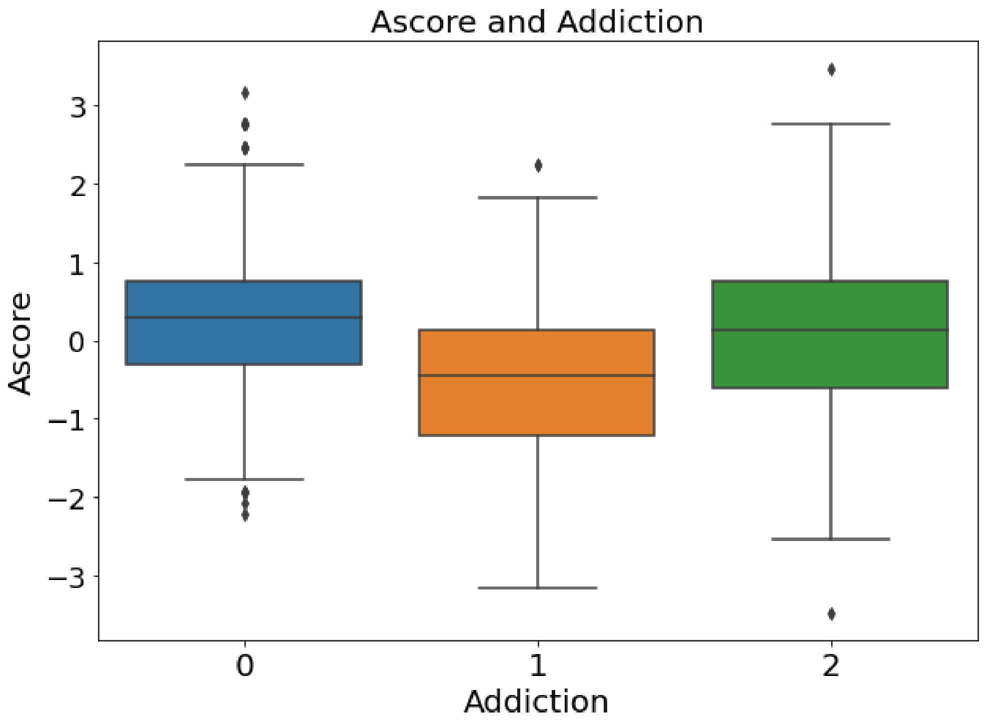
- Escore: Extraversion Score
- Nscore: Neurortism Score
- Oscore: Open minded Score



NEO PI-R: Personality Survey

- Cscore: Conscientiousness Score
- Ascore: Agreeableness Score





NEO PI-R: Conclusion

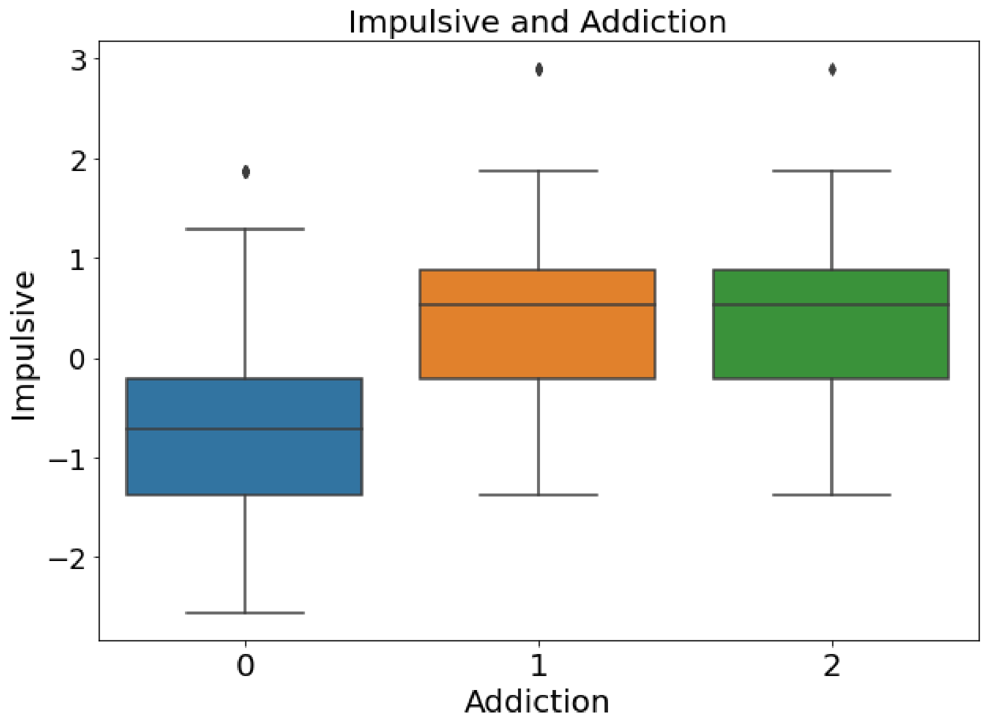
All the scores seem to have an impact on clustering

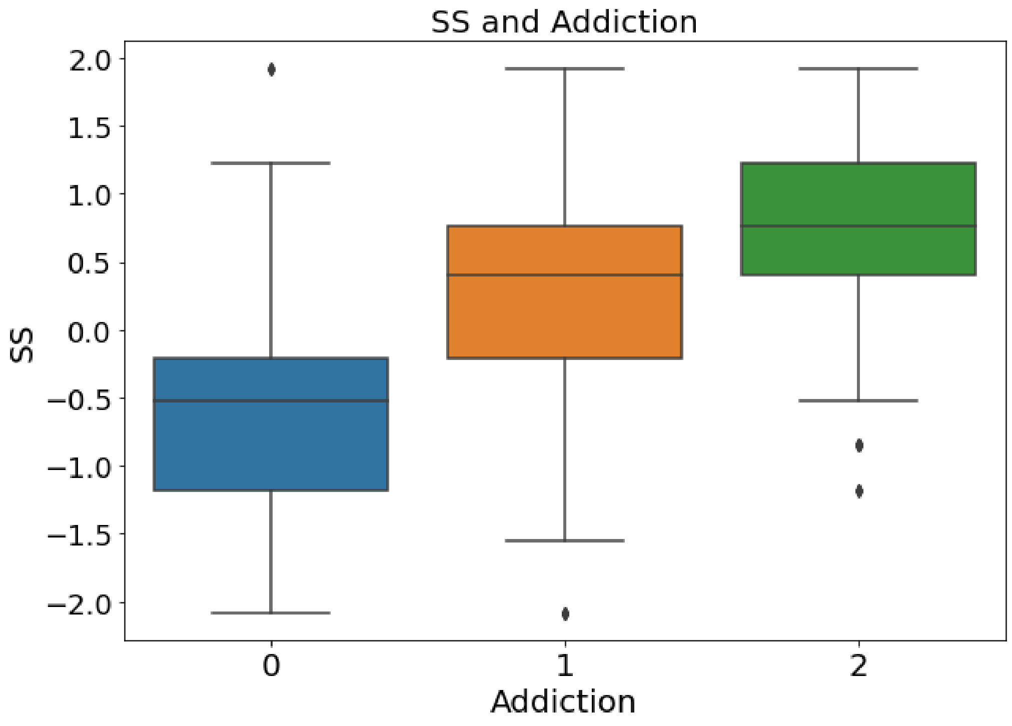
Group 0 is constant and probably show the score of non-user people

Group 1 and 2 show many variations

Impulsiveness and Sensation measure

- Impulsive is Impulsivness by BIS-11 survey
- SS is Sensation by ImpSS survey





Sensation measures: Conclusion

Impulsiveness define the diference between class 0 and the others

SS better describe the 3 groups

Assumption: SS will have a better impact on our prediction model

Drugs

- Analyze of the groups distributions
- We looked at drug type*(link)
- 0 = Popular Drugs (Alcohol, Chocolate …)
- 1 & 2 = Illegal/Dangerous Drugs

Drug	Type	Dependency	Cluster group more frequent	Obs.
Semer	Fictional Drug	NaN	NaN	Fictional Drug where 99% answer they would'nt try
Alcohol	Not specified as a Drug	High Risk	0	Well spread in the population
Amphet	Stimulating	High Risk	2	Group 1 is close to group 2
Amyl	Depressant	Low Risk	2	Group 1 is close to group 2
Benzos	Neuroleptics/Depressant	Medium Risk	1	Group 1 is well defined
Caff	Stimulating	NaN	0	Well spread in population
Cannabis	Various	Medium Risk	1-2	1 and 2 similar
Chocolate	Not specified	NaN	0	Well spread in population
Cocaine	Stimulating	High Risk	1	Group 1 well defined

*https://fr.wikipedia.org/wiki/Classification_des_psychotropes

Drugs

- 19 Drugs were analyzed
- First analyse the groups distributions
- We looked at drug type*(link)

Drug	Type	Dependency	Cluster group more frequent	Obs.
Crack	Stimulating	High Risk	1	0 group is not addicted, 1 is the most
Ecstasy	Stimulating/Neuroleptics/Halluconigenics	Low Risk	2	group 2 is a little more than group 1
Heroin	Depressant	High Risk	1	group 1 is a little more
Ketamine	Depressant/Halluconigenic	High Risk	2	group 2 more
Legalh	Drug substitute	Medium Risk	2	group 0 low, group 1 and 2 similar
LSD	Stimulating/Halluconigenics	Low Risk	2	group 2 the most
Meth	Stimulating	High Risk	1	group 1 the most
Mushrooms	Halluconigenic	Low Risk	2	group 2 the most
Nicotine	Neuroleptics/Stimulating/Depressant	High Risk	1	group 1

*https://fr.wikipedia.org/wiki/Classification_des_psychotropes

Univariate Analysis: Conclusion

Group 0 explain non-user/non addicted people.

Group 1 and Group 2 are'nt well defined with this first analysis, but we can make the hypothesis that group 1 the use of more violent drugs than group 2

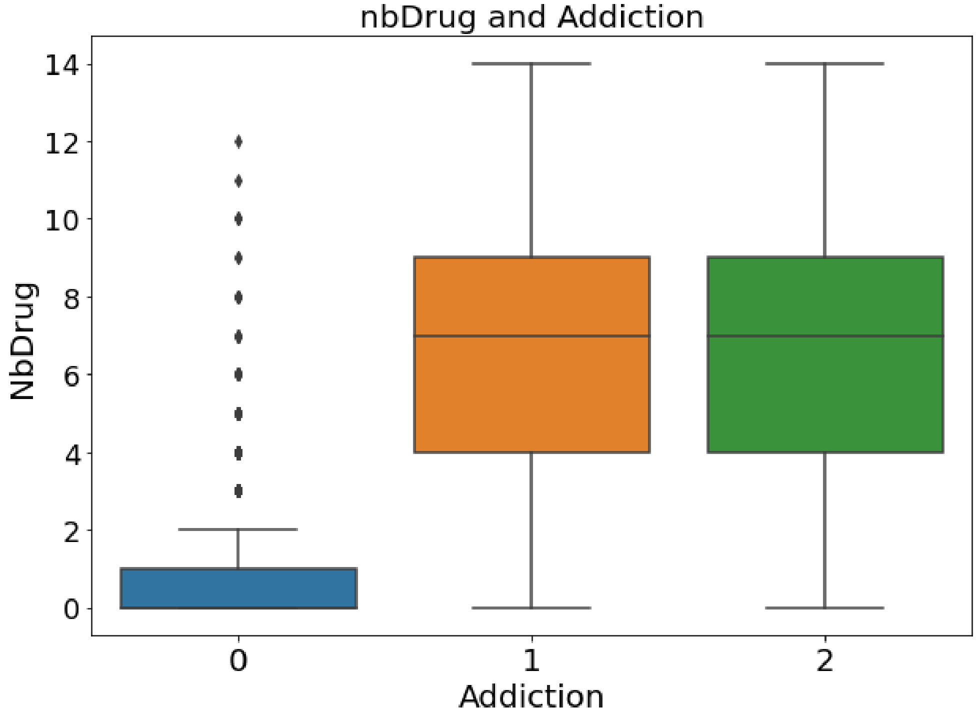
Semer, Alcohol, Caffeine, Chocolate and Nicotine have to be removed because of beeing too popular or not popular enough.

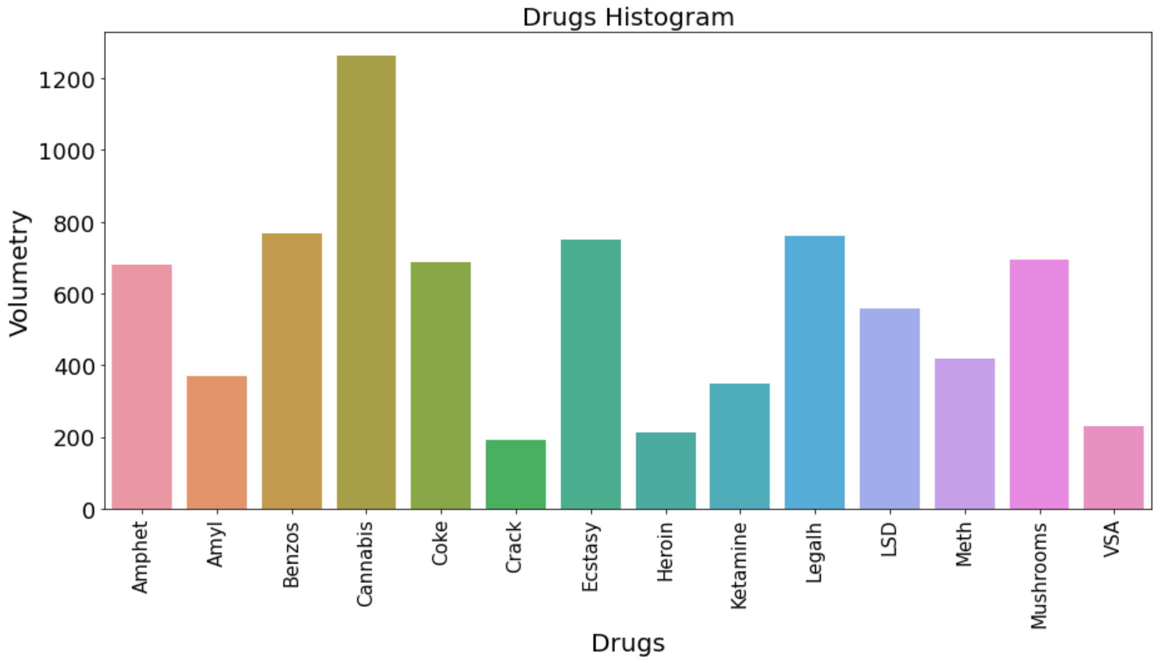
Ethnicity can be dropped because it doesn't have any effect.

Multivariate Analysis

Drugs

- Cannabis is the most popular one
- Confirmation that 0 represents the non-user population
- 1 & 2 aren't differentiated by the quantity



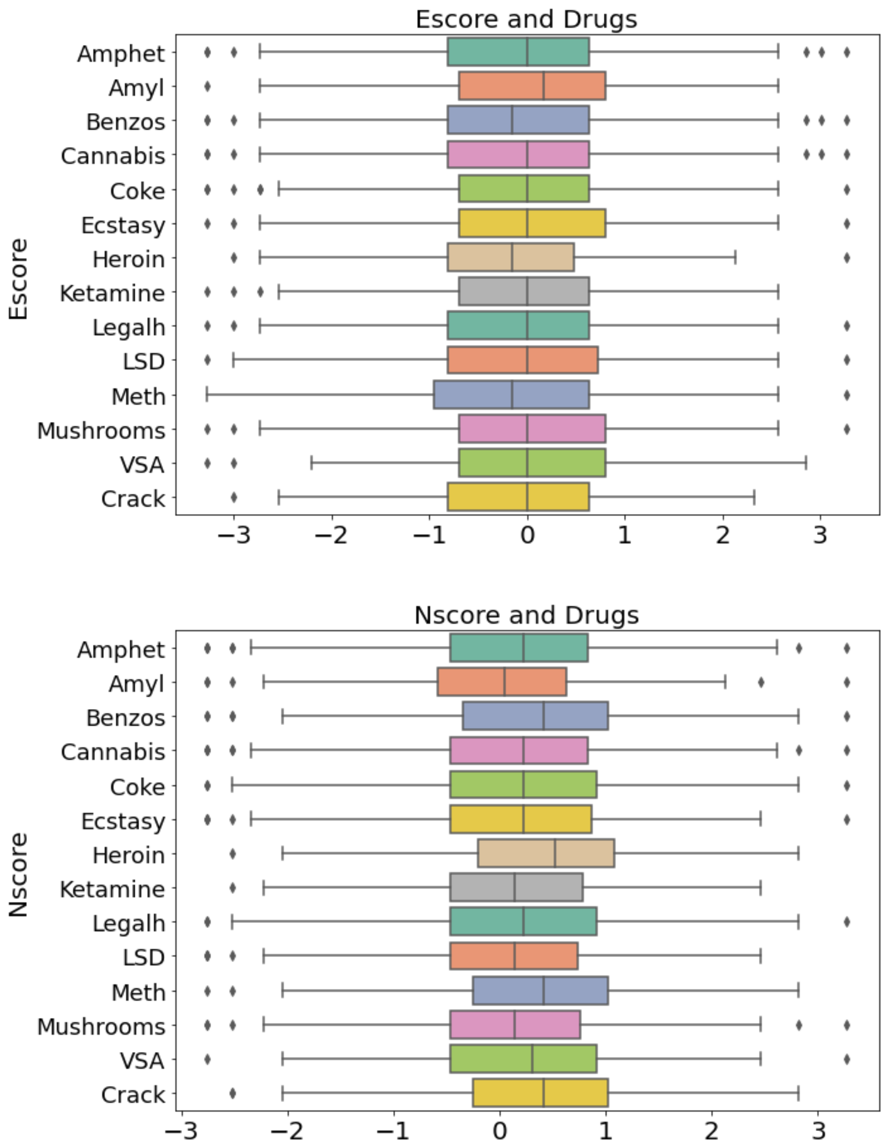


Drugs

- Cannabis is the most popular one
- Confirmation that 0 represents the non-user population
- 1 & 2 aren't differentiated by the quantity
- Cannabis is the most consumed drug

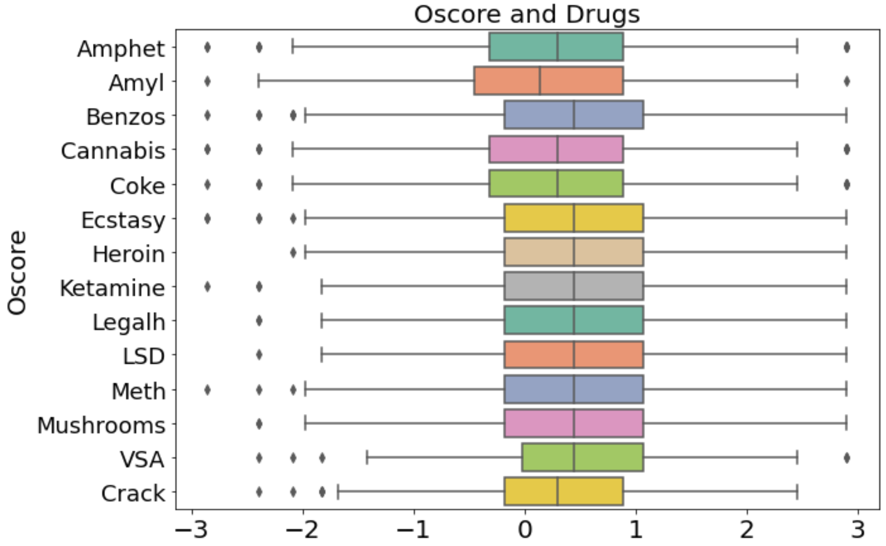
Scores & Drugs

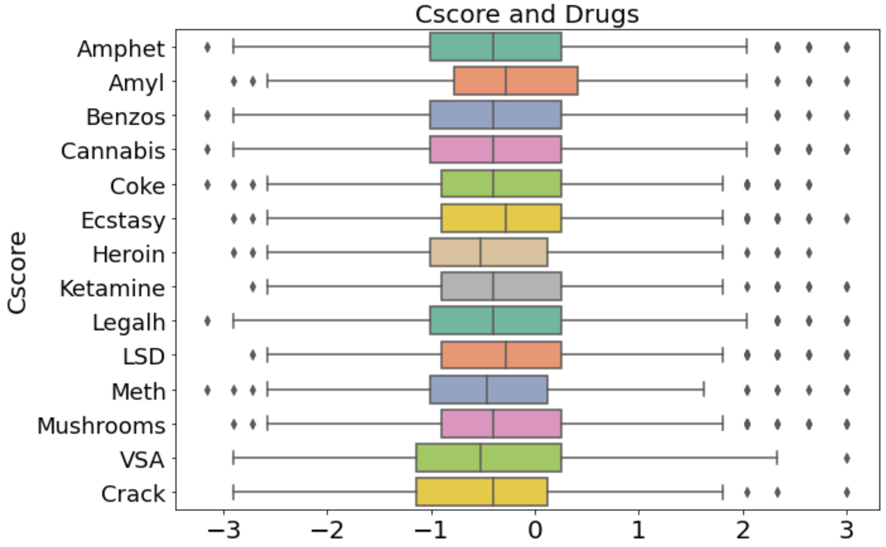
- **Nscore**:
  - High positive scores show group 1 drugs:
    - Benzos, Meth, Heroin
  - Low positive scores show group 2 drugs:
    - Amyl, Mushrooms, Ketamine
- **Escore:**
  - High positive scores show group 2 drugs:
    - Amyl
  - Less positive scores show group 1 drugs:
    - Benzos, Heroin, Meth



Scores & Drugs

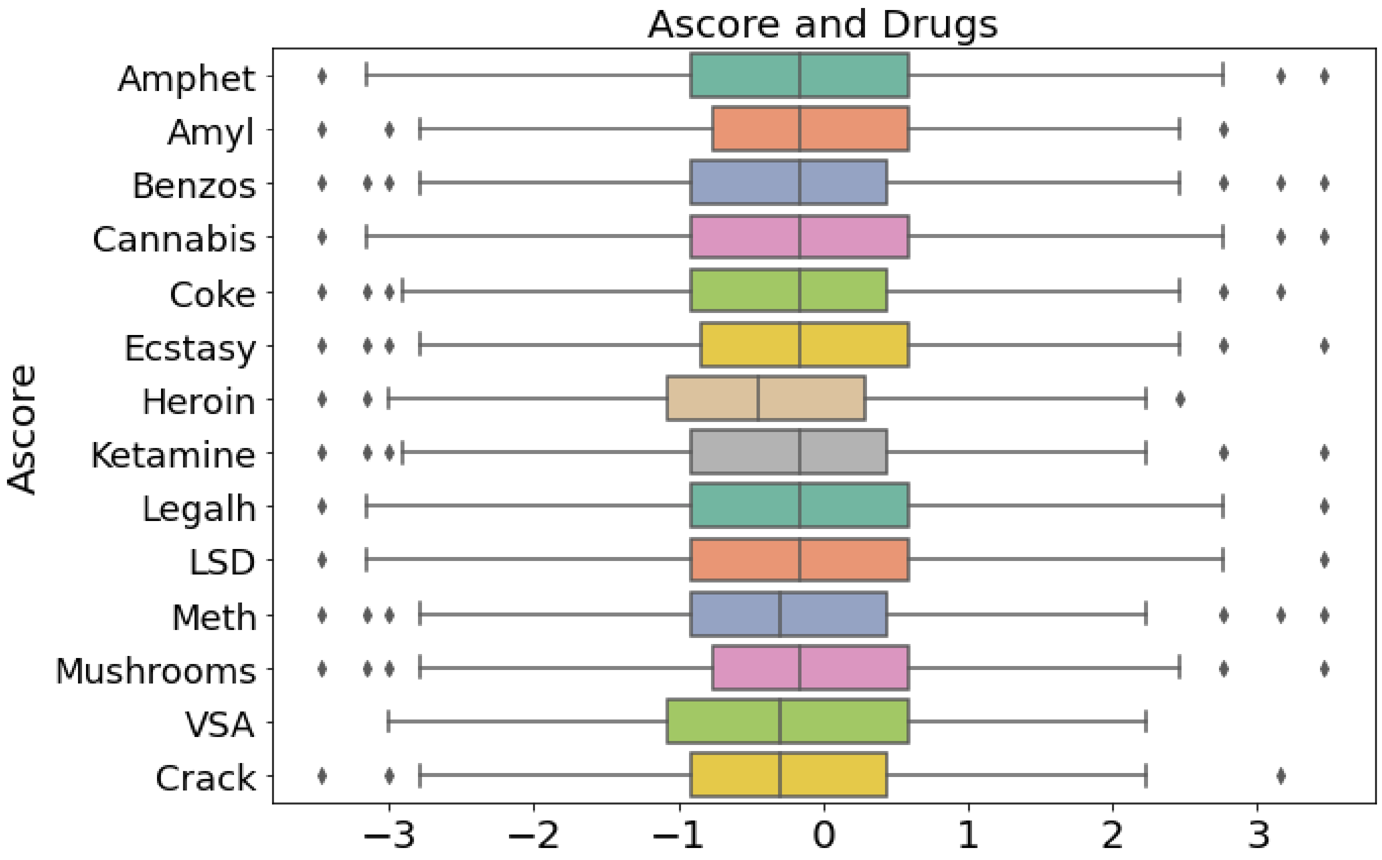
- **Oscore**:
  - Similar for each drugs
  - This score is highly present with the use of any drugs.
- **Cscore:**
  - High negative scores show group 1 drugs:
    - Meth, Heroin, VSA
  - Less negative scores show group 2 drugs:
    - Amyl





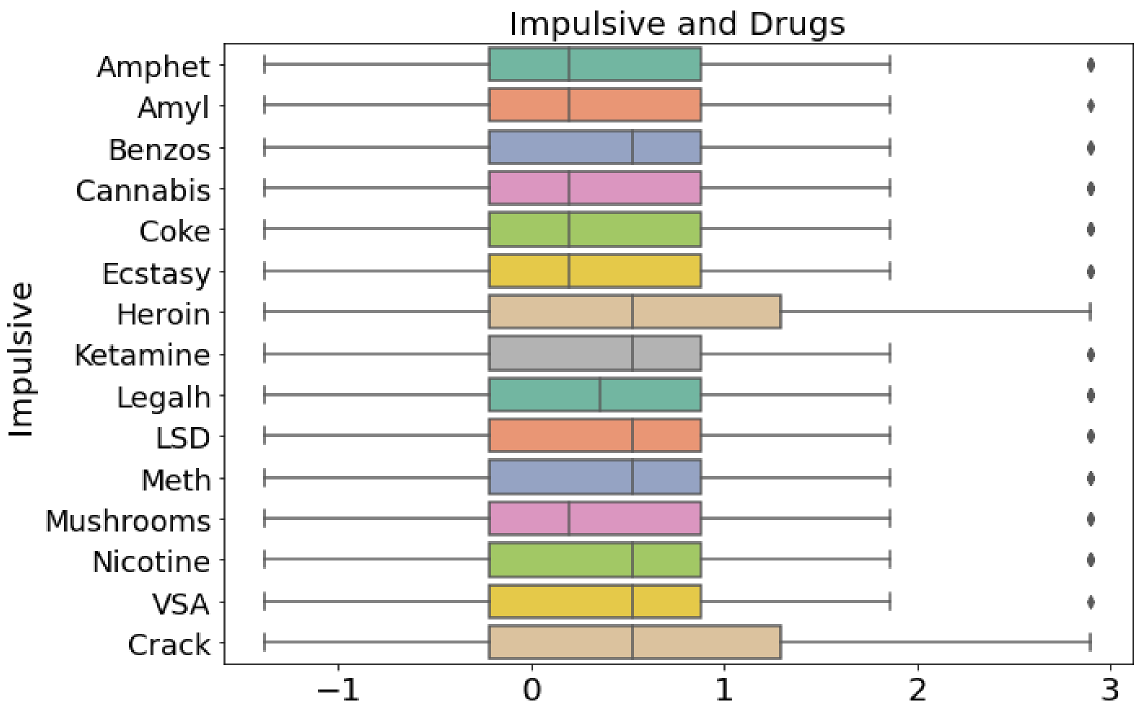
Scores & Drugs

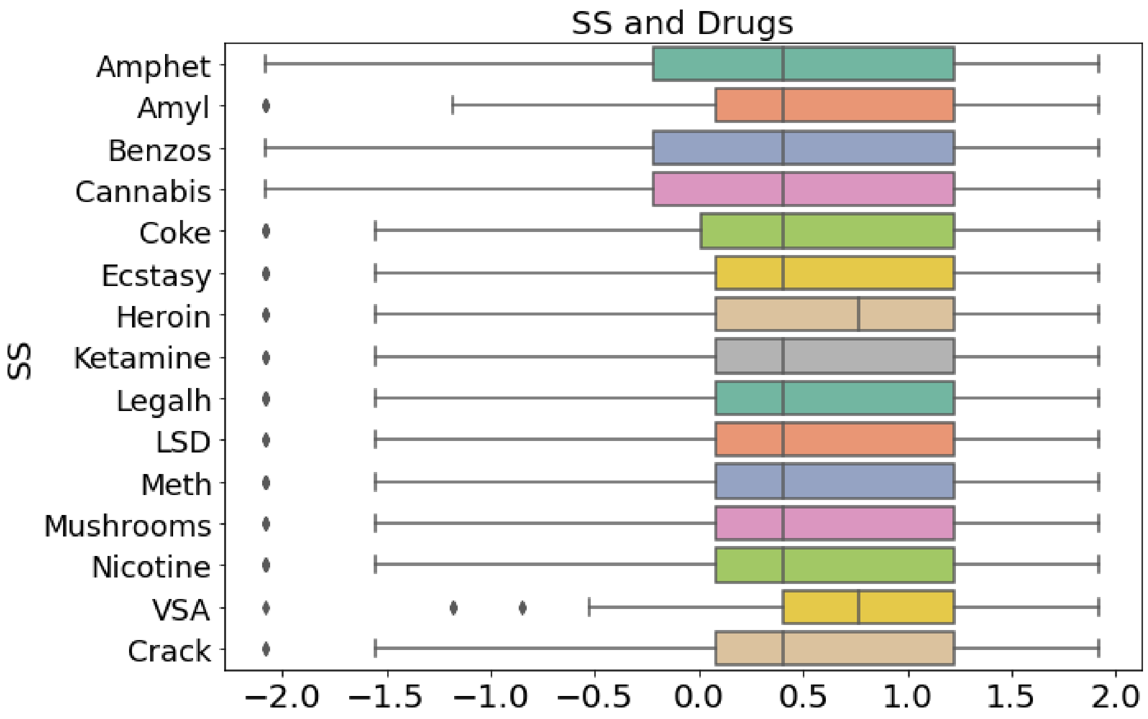
- High negative AScore show group 1 people:
  - Heroin, Crack, Meth, VSA
- As interpreted in univariate analysis, drugs from group 2 aren't affected by Ascore, the scores are similar to group 0 population



Impulsivity and Sensation vs Drugs

- Impulsivity and SS are more extreme for group 1 drugs
  - Heroin, Coke, VSA





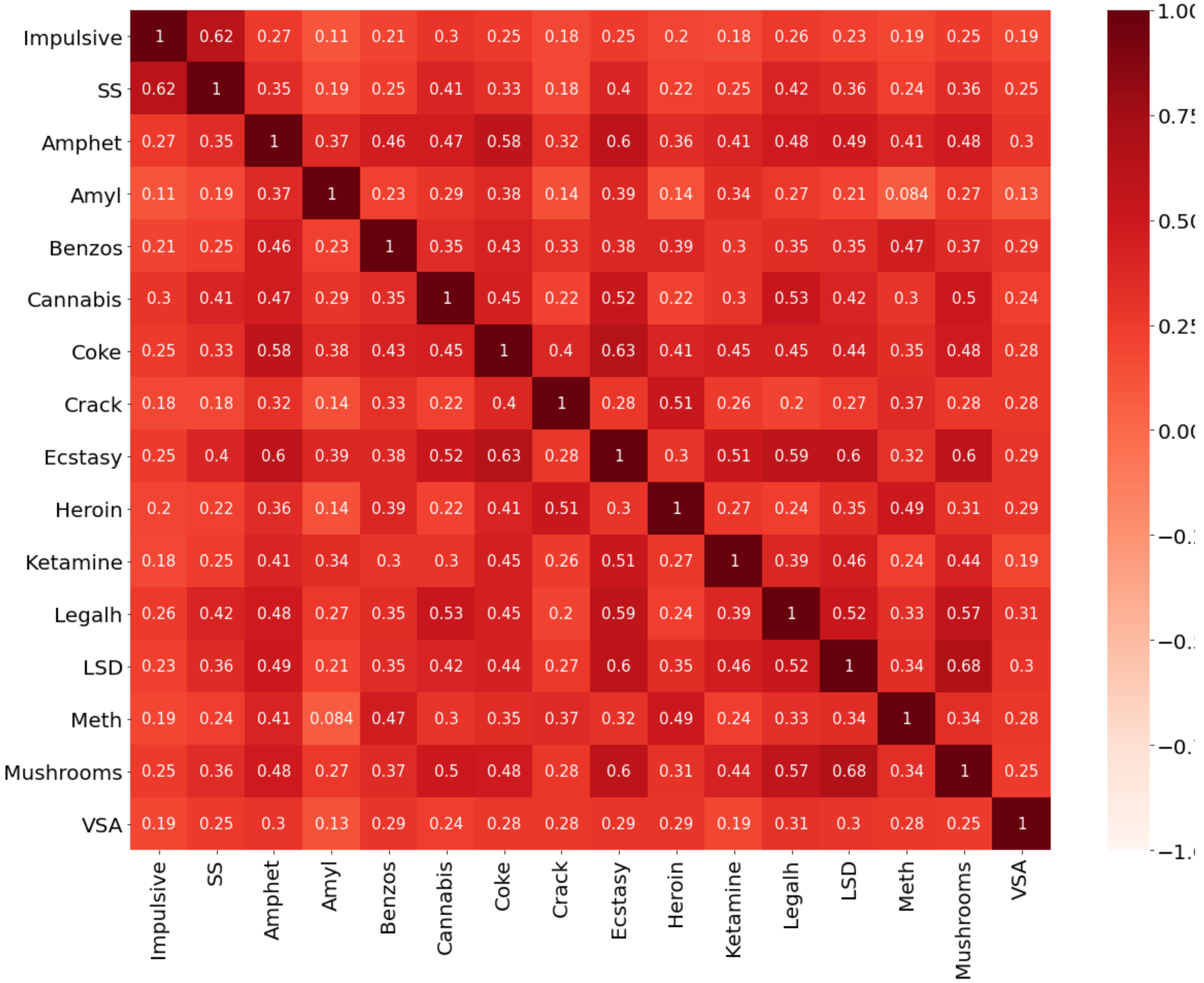
Multivariate Analysis: Conclusion

- We can confirm that group 0 represents non-user people
- Group 1 and group 2 represent user people
- What differentiate group 1 and 2 is the psychologics Scores
- Group 2 is very close to normal scores (Group 0), so we can consider group 2 as an addiction without or with low violent psychologics effects
- Group 1 represents people with violent psychologics effects

Correlation Analysis

Heatmap:

- The scores aren't correlated (for better visualization we removed them)
- SS and impulsive are highly correlated
    - We choose SS because it is highly correlated to the target (univariate analysis)
- Some Drugs are correlated:
    - Heroin and Crack (Crack is made of heroin)
    - Mushrooms and Ecstasy are correlated with many variables



Correlation: Final Dataset to model

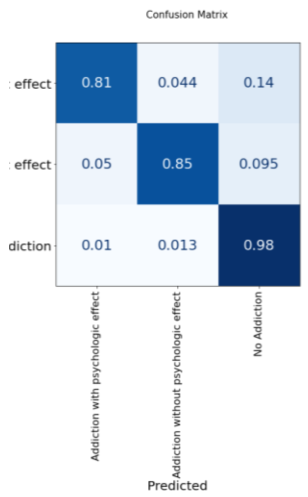
We will keep the following features to model predictions

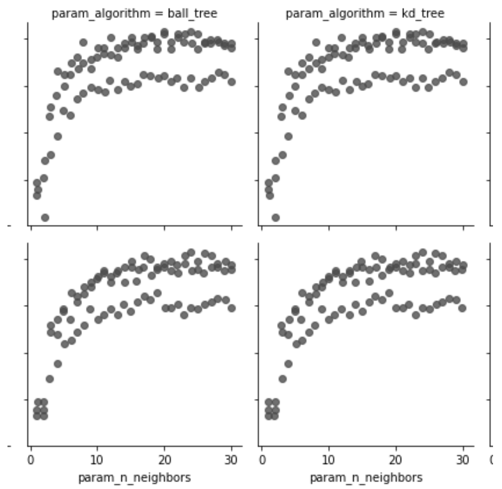
20 variables: 10 float (score and age) and 10 categorical (drugs)

Modelisation

K-NN

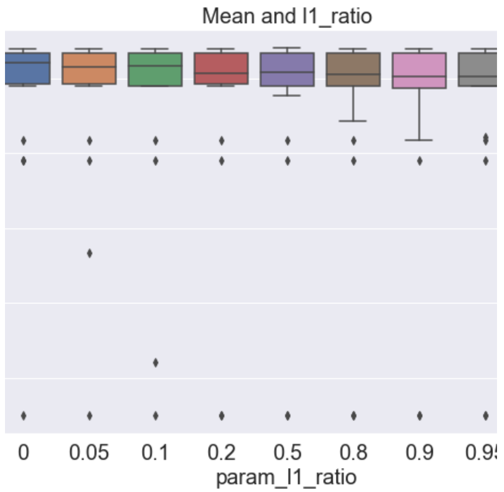
- Train best score: 0.902
- Test score: 0. 898
- Cross-validation: Kfold
- Algorithm and wheights hyperparameters don’t have a relevant impact
- K neighbors converge after k = 20
- No addiction is well predicted

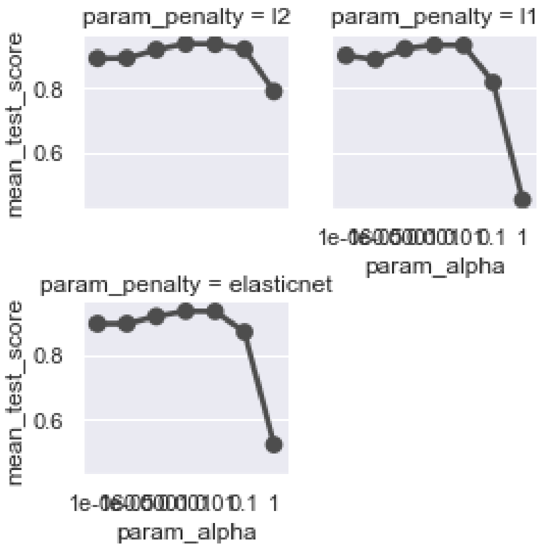


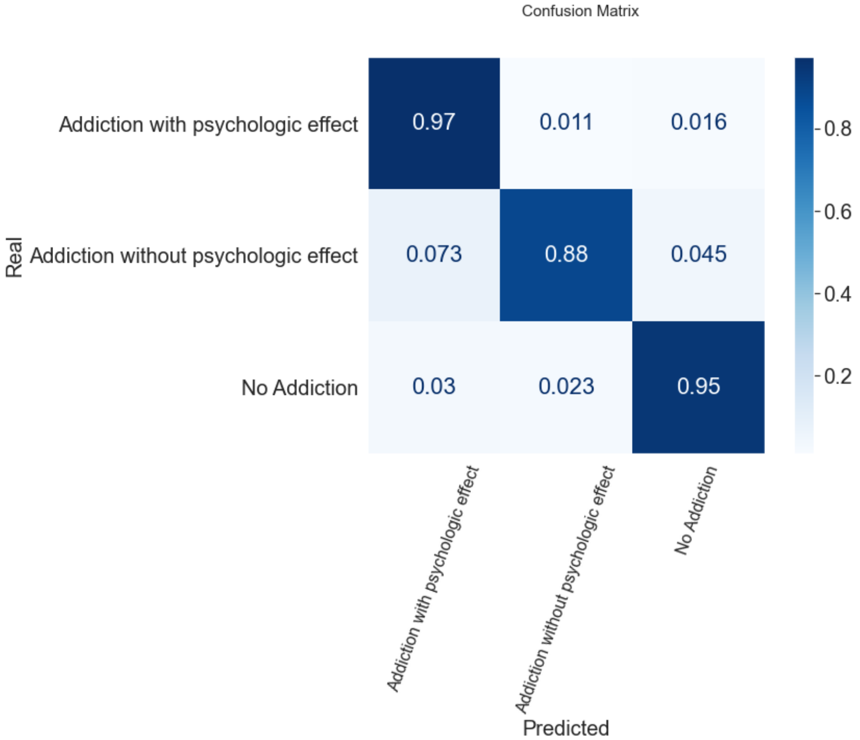


SGDClassifier

- Train best score: 0.94
- Test score: 0.936
- Cross-validation: KFOLD
- L1 ratio have no impact
- Penalty have few change if alpha < 1
- Alpha is the key parameter
- Great prediction for addiction with psychologic effect

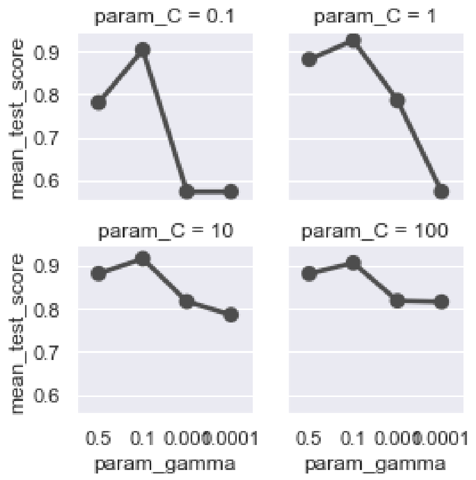


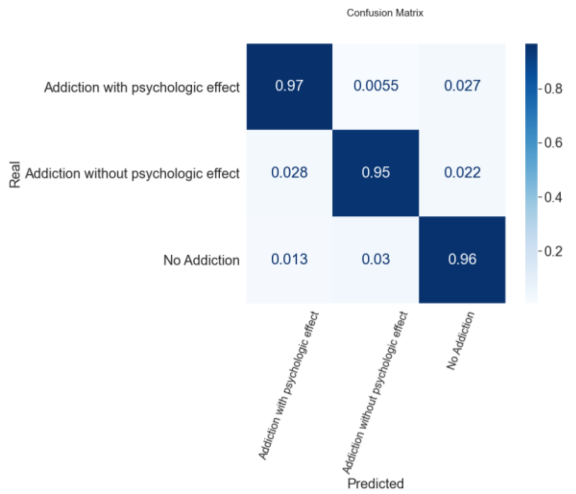


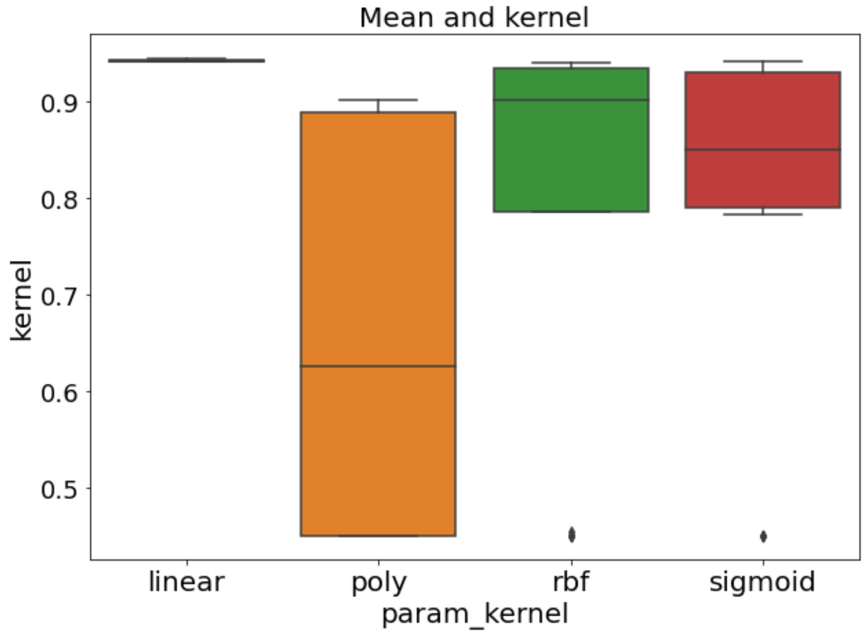


SVC

- Train best score: 0.945
- Test score: 0.957
- Cross-validation: KFOLD
- the algorithm tends to work better with gamma = 0.1, C = 1 or 10
- linear kernel seems to have a better behavior
- it cannot be improved a lot and the best score will not have big change with different parameters
- Good overall prediction





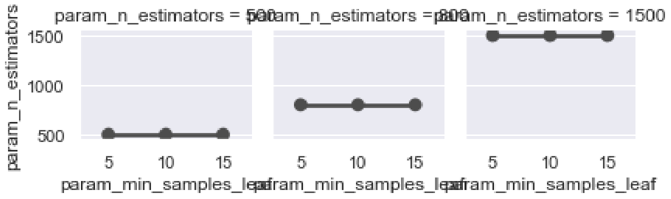


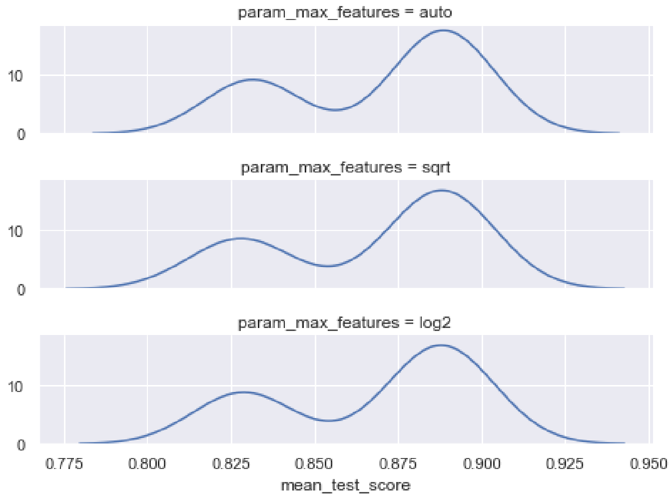
Decision Tree Algorithms

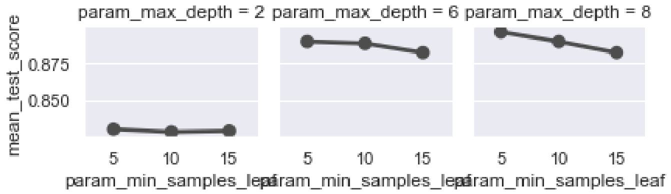
- Random Forest
- Gradient Boosting
- XGBoost
- AdaBoost

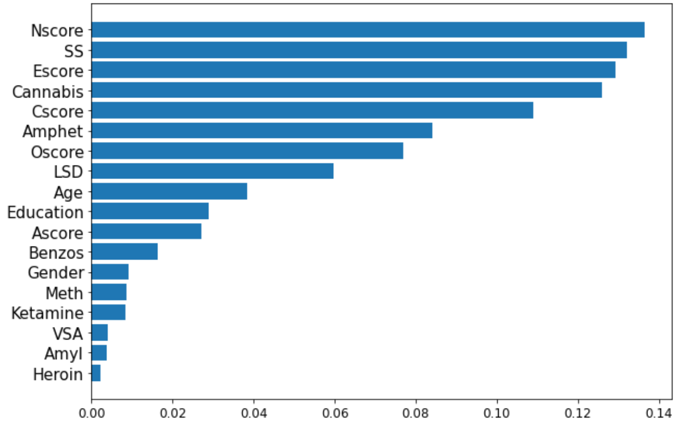
Random Forest

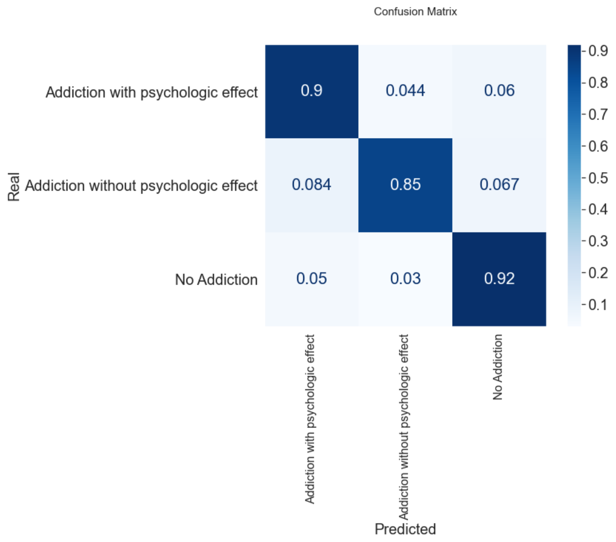
- After a first train we get the feature importance and we decide to remove VSA, Amyl, Heroin
- Train best score: 0.897
- Test score: 0.893
- Accuracy increase when:
    - max depth increase
    - Min sample leaf decrease
    - N_estimator increase
- Better prediction for no addicted population





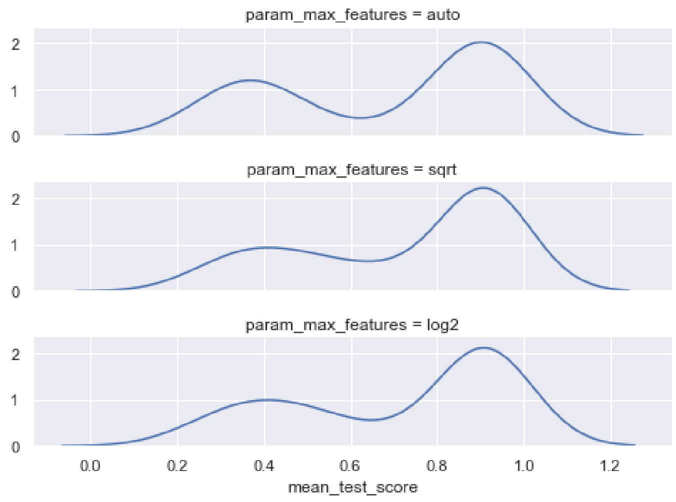


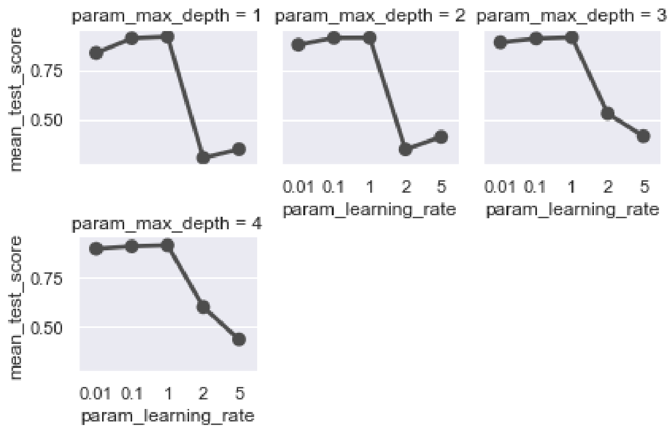


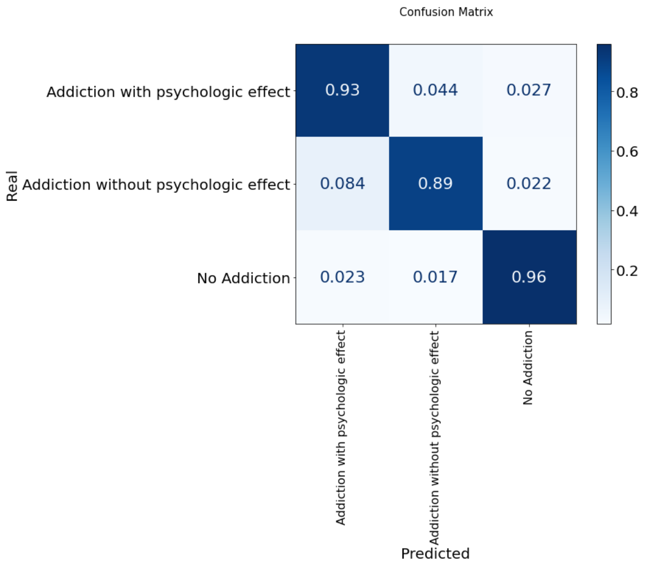


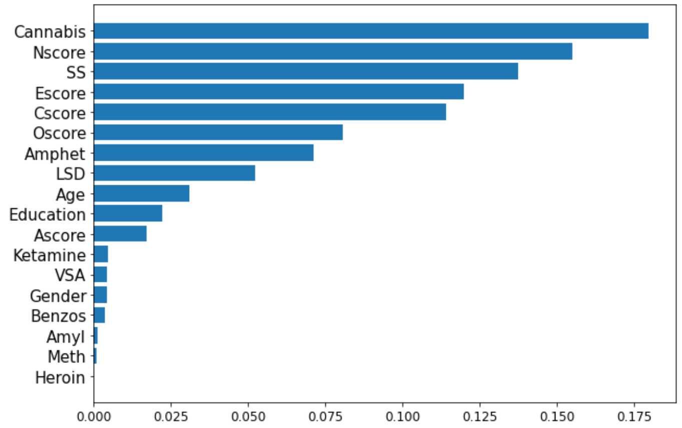
Gradient Boosting

- After a first train we get the feature importance and we decide to remove Meth, Amyl, Heroin
- Train best score: 0.929
- Test score: 0.933
- Max features doesn't see to have an important effect
- Learning rate is the key: ideal learning rate is 1
- Good prediction for no addicted population



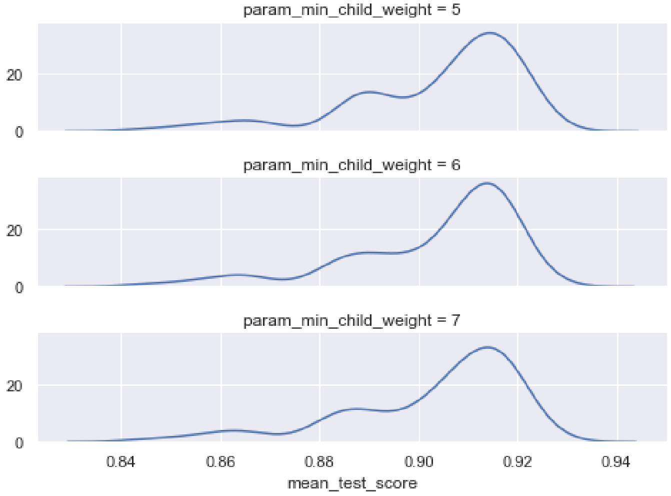


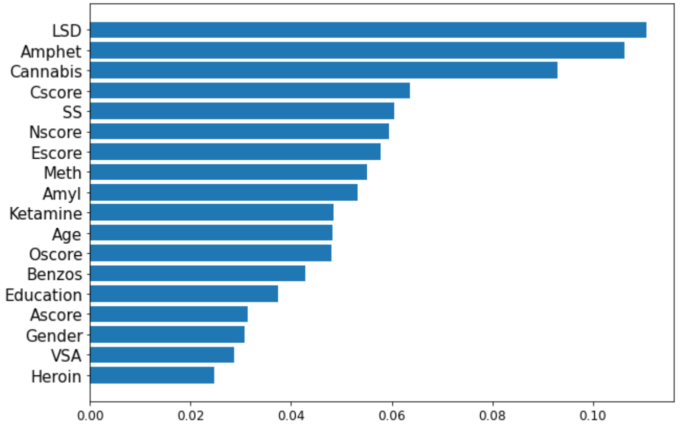


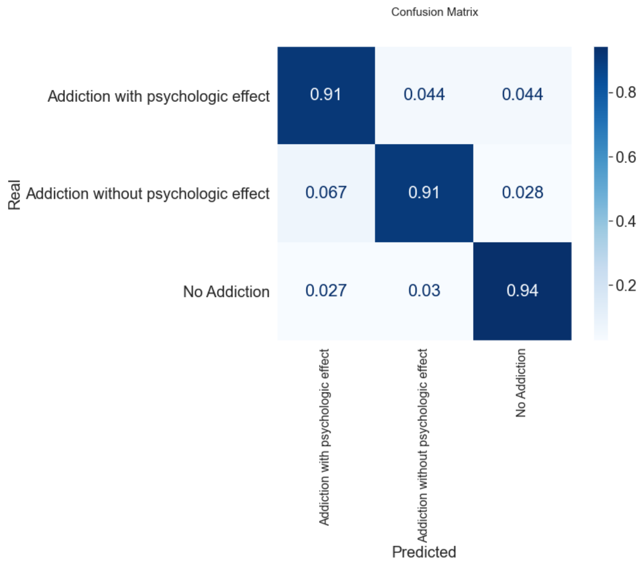


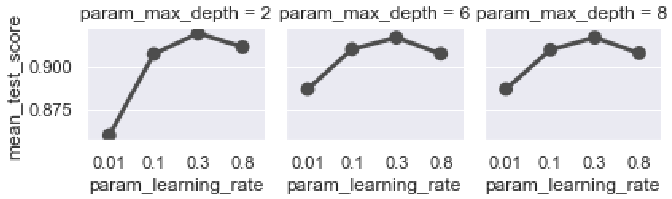
XGBoost

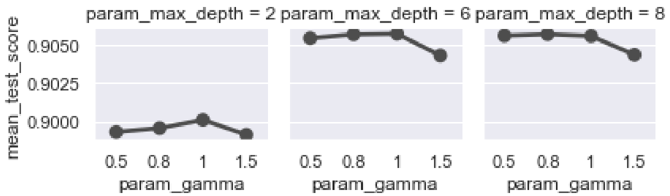
- After a first train we get the feature importance and we decide to remove VSA, Gender, Heroin
- Train Best score 0.933
- Test score 0.919
- 2 parameters have a real impact
  - Learning rate: ideal = 0.3
  - gamma: ideal = 1
- Great predictions for no addiction population





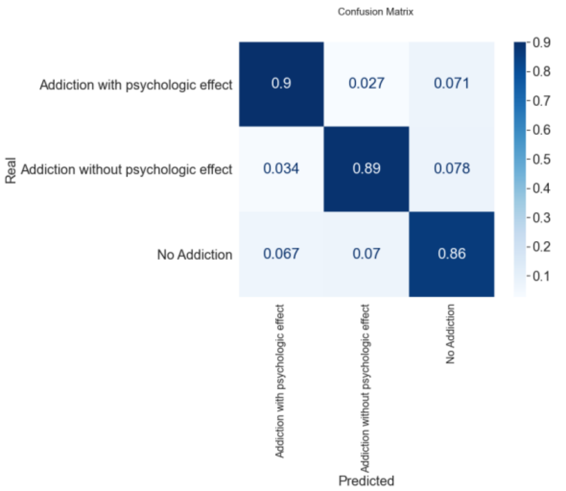


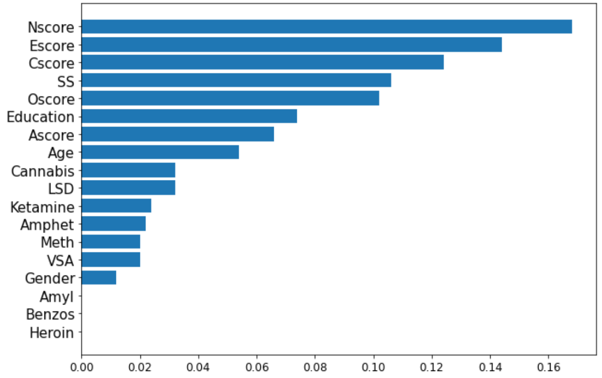


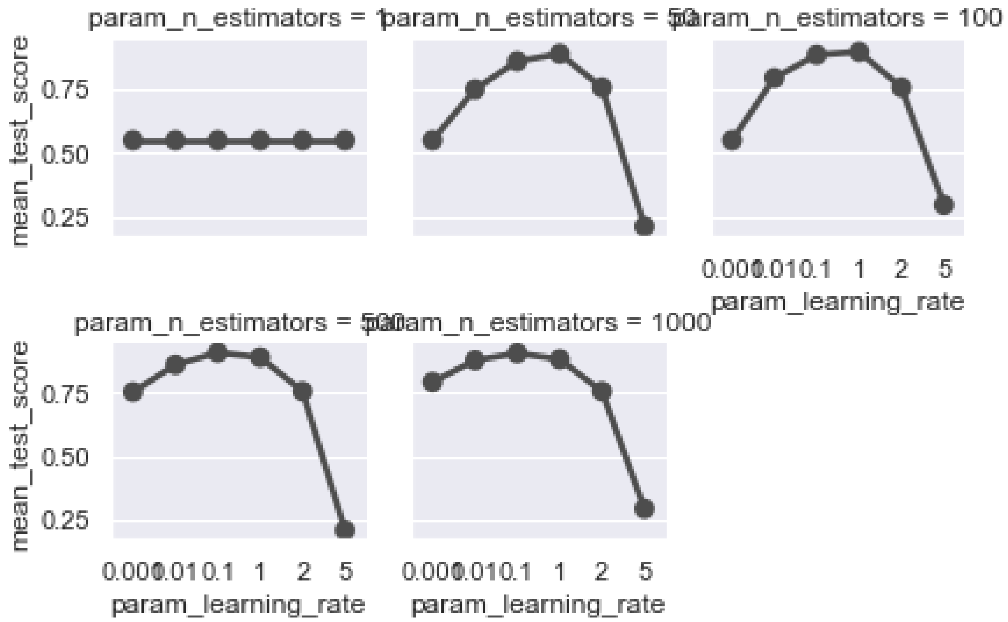


AdaBoost

- After a first train we get the feature importance and we decide to remove Meth, Benzos, Heroin
- Train best score: 0.90
- Test score: 0.915
- N-estimator doesn't seem to have an effect
- Learning rate is the key hyperparameter and the ideals are 1 or 0.1
- Good prediction for addicted with effect







Model Conclusion

- SVC seems to be the better model.
  - A good SVC development is proof of a good initial clustering.
  - Our classes are globally well separated, and the prediction works well
- Decision Trees algorithms have good results
  - Heroin, VSA are often useless
- Maybe a research with medical areas to penalize some features or oversampling too avoid unbalance problem will improve the model

Model	Train	Test	Better prediction group
K-NN	0.902	0.898	No addiction
SGDClassifier	0.94	0.936	Addiction with psychologic effect
SVC	0.945	0.957	Addiction with psychologic effect
Random Forest	0.897	0.892	No addiction
Gradient Boosting	0.929	0.933	No addiction
XGBoost	0.933	0.929	No addiction
AdaBoost	0.90	0.915	Addiction with psychologic effect

Project Conclusion

SVC will be used in the API to make predictions

This model can be used to classify patient in a medical center.

Very important to get great « no addiction » predictions in order to give medicine to the right person and not make mistake.

Some features to improve the model

Patient Social Conditions

Historical data about the patient

API

FLASK

Namespaces

dataset	Dataset related endpoints
filtering	Filtering phase in which we have the plots which we use in order to do the pre processing
model	Model related endpoints

Processing plots

- There are 21 endpoints which are returning the most important plots.

filtering Filtering phase in which we have the plots which we use in order to do the pre processing

GET	`/filtering/boxplot/addiction/age`	Get the boxplot for age in function of addiction
GET	`/filtering/boxplot/addiction/ascore`	Get the boxplot for ascore in function of addiction
GET	`/filtering/boxplot/addiction/cscore`	Get the boxplot for cscore in function of addiction
GET	`/filtering/boxplot/addiction/education`	Get the boxplot for education in function of addiction
GET	`/filtering/boxplot/addiction/escore`	Get the boxplot for escore in function of addiction

K-NN: First Model

- K-NN related endpoints

model

Model related endpoints

POST `/model/knn` Predict the class of the consumer with knn method

GET `/model/knn/confusion_matrix` Get the Confusion Matrix for the knn method

GET `/model/knn/elbow` Get elbow method graph for the knn clustering

SVC: second model

- SVC related endpoints

POST `/model/svc` Predict the class of the consumer with svc method

GET `/model/svc/confusion_matrix` Get the Confusion Matrix for the svc method