ADDICTION TO DRUGS

DESEURE--CHARRON FLAVIEN & DEMOUGE BENJAMIN

INTRODUCTION

I – DATA VISUALISATION

Global comprehension

Exploration

Creation of different visuals

Relationships

II – PREPROCESSING

Target selection

Target encoding

Feature selection

Feature encoding

III – MODELING

Metrics

Phases

- Initial classes
- New classes

Tuning hyperparameters

IV – FINAL MODELS

Comparison

Selection

PRESENTATION OF OUR DATASET

- Our dataset came from this site, It is the result of a study from 2011 to 2012 on 1885 adult English speakers.

- the study allows us to have a better understanding of the relationship between categories of people and their consumption of any king of drugs.

- Seven features were related to three different personality test, NEO-FFI-R, BIS-11 and ImpSS.

Category of drugs	Drugs (19)
Common	Caffe and Chocolat
Substances diverted into drugs	Volatile substance Abuse, Benzodiazépine, Ketamine, Methadone and Mushrooms
Legal	Alcohol, LegalH and Nicotine
Illegal	Amphetamines, Nitrite d'Amyle, Cannabis, Cocaïne, Crack, Esctasy, Heroin and LSD
Fictional	Semer

PRESENTATION OF OUR DATASET

- The data set contained information on the consumption of 18 central nervous system psychoactive drugs.

- For each drug, the last consumption frequency is indicated as in the following table:

Frequency of consumption
Never Used
Used over a Decade Ago
Used in Last Decade
Used in Last Year
Used in Last Month
Used in Last Week
Used in Last Day

EXPLANATION OF THE 3 PERSONNALITY TESTS

NEO-FFI-R: The Big Five personality test measures the five personality factors that psychologists have determined are core to our personality makeup.

- Nscore: Neuroticism - How sensitive a person is to stress and negative emotional triggers.
- Escore: Extraversion - How much a person is energized by the outside world.
- Oscore: Openness - How open a person is to new ideas and experiences.
- Ascore: Agreeableness - How much a person puts others' interests and needs ahead of their own.
- Cscore: Conscientiousness - How goal-directed, persistent, and organized a person is.

BIS11: Barratt Impulsiveness Scale (BIS-11) is a questionnaire designed to assess the personality/behavioral construct of impulsiveness

ImpSS: The ImpSS scale is a 19 questions true-false scale assessing various personality characteristics and behaviors related to impulsivity and sensation seeking, and it is scored by summing the items that are consistent with impulsivity or sensation seeking. Thus, scores for this scale range from 0 to 19.

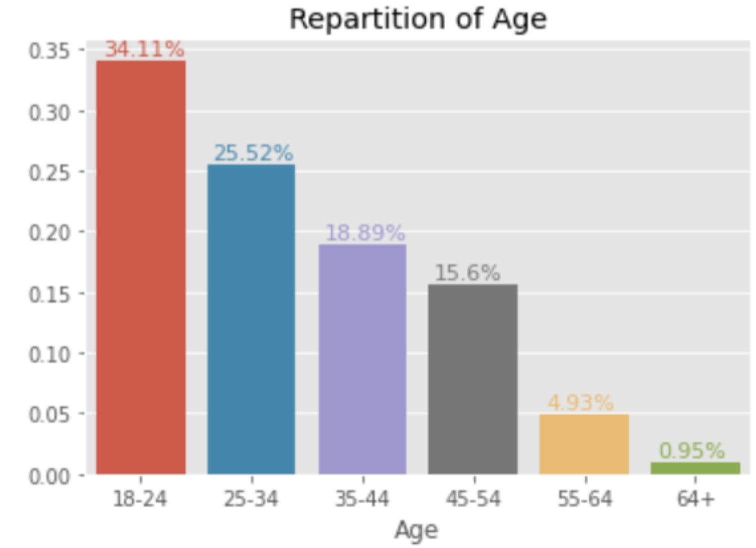
How can we model the risk of addiction to a drug based on personality and demographic data?

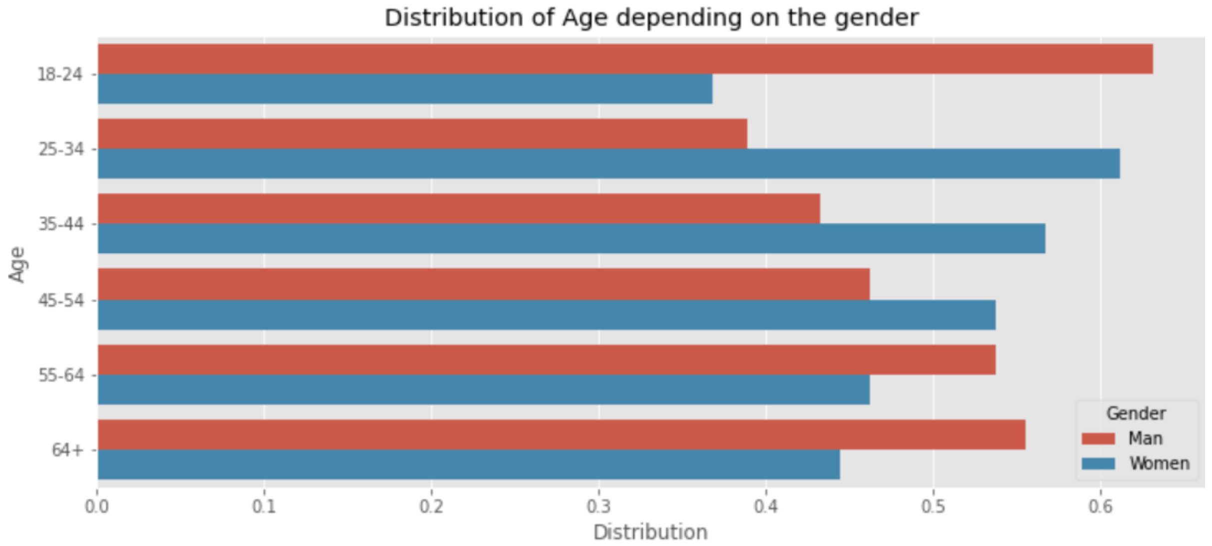
I – DATA VISUALISATION

- Visual linked to the repartition of the population by age

- We can see that most of our population is young

- Also, man over represent women in the 18/24-year-old category and the opposite on the 25/34-year-old category



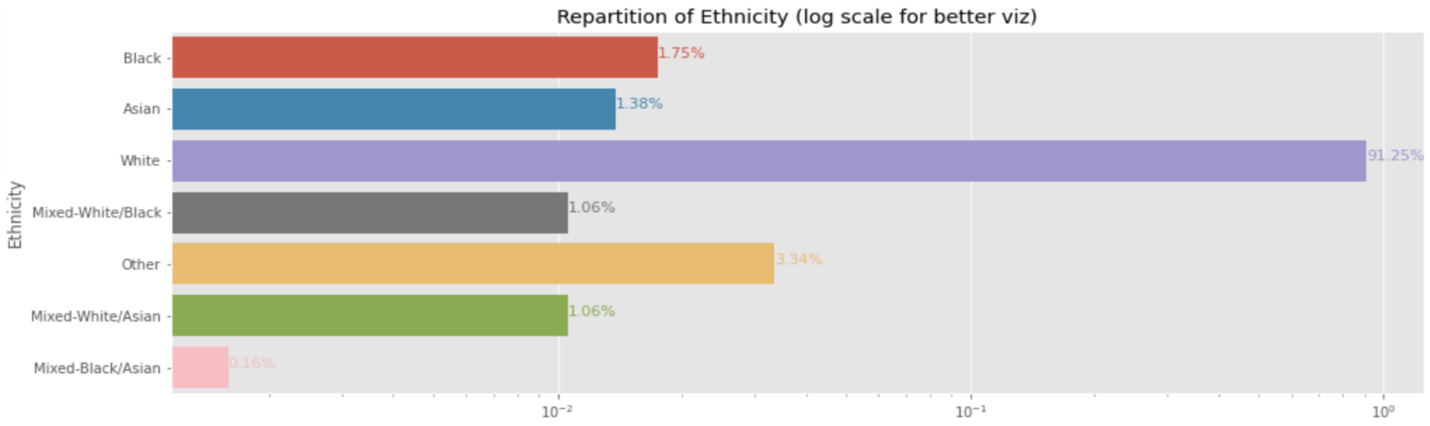


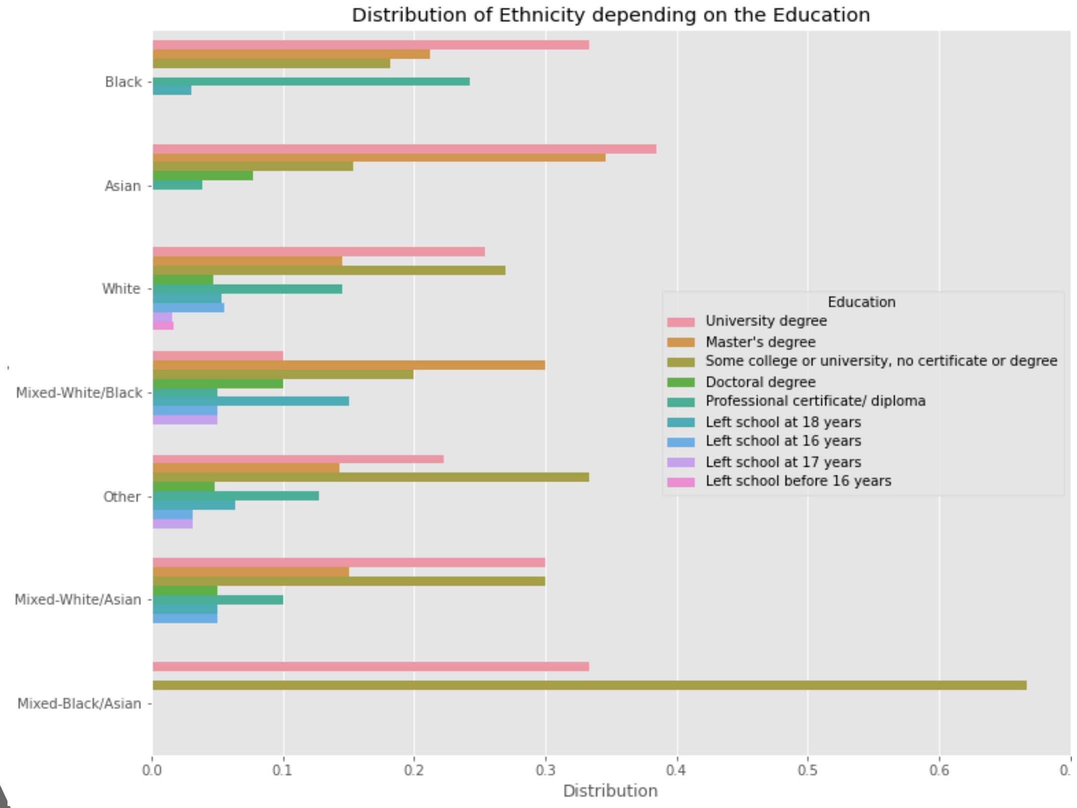
I – DATA VISUALISATION

- Visual linked to the repartition of the population by Ethnicity and Education

- The group of white people represent more that 90% of our population, Meaning that our prediction model would not have enough data for the other ethnicity.

- There is no correlation between Ethnicity and Education, however we could notice that our population is more educated-people than left-school-before-19-people.

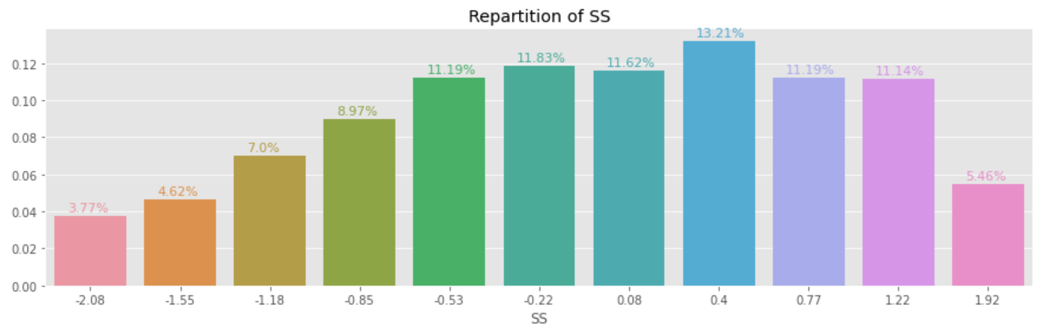


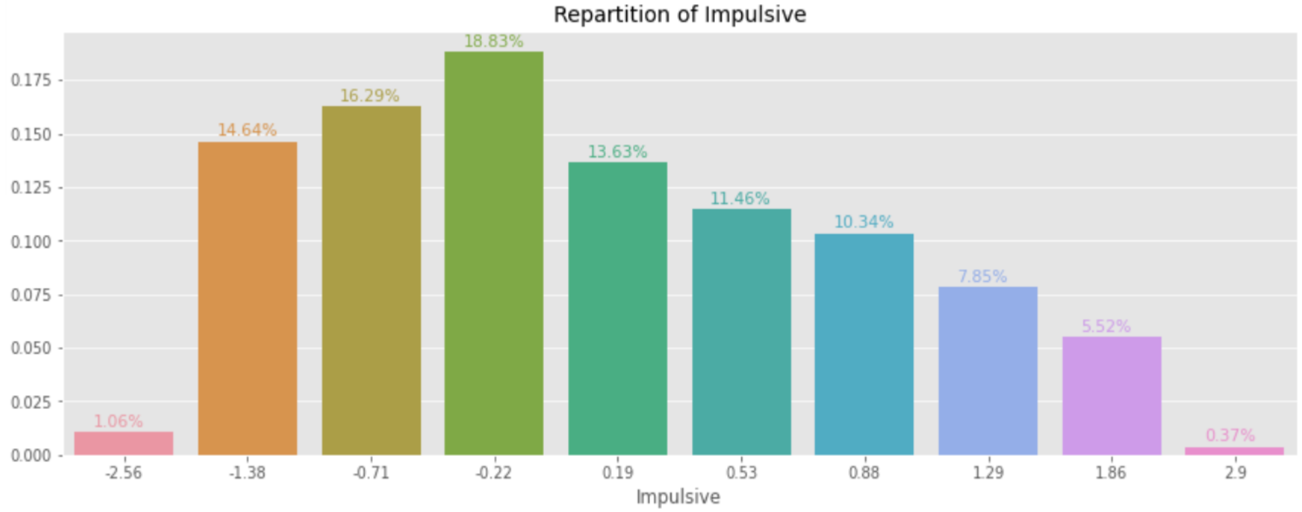


I – DATA VISUALISATION

- Visual linked to the repartition of the population by the score they got in the ImpSS and BIS-11 test

- We see that both of our repartition is close to a gaussian distribution.
- Therefore, we can conclude that our population represents properly the entire population on a personal level as it is related to personality test



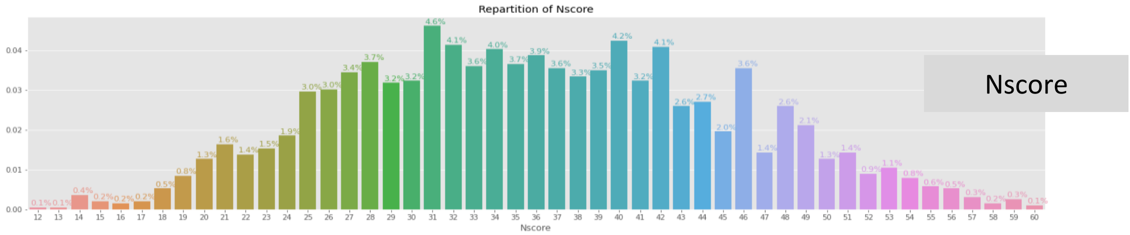


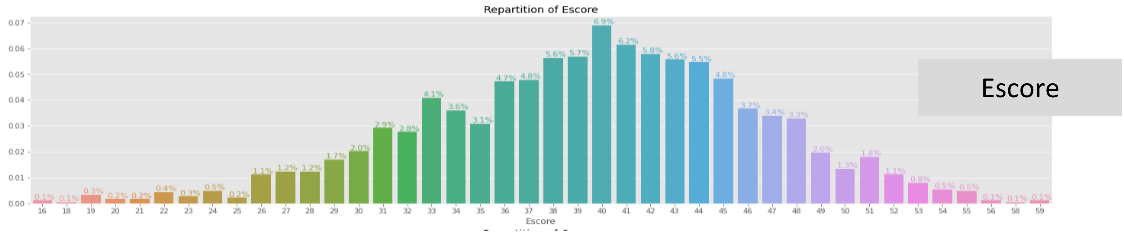
I – DATA VISUALISATION

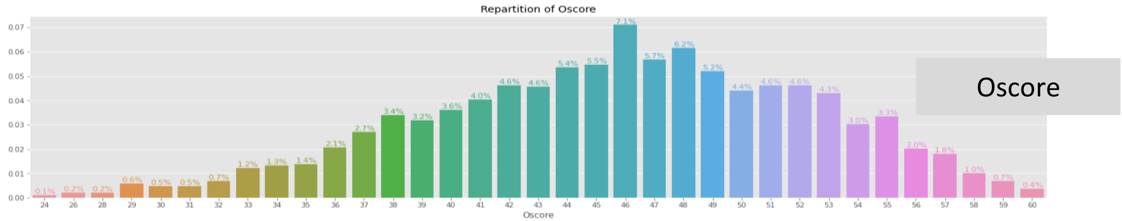
- Visual linked to the repartition of the population by the score they got in the NEO-FFI-R test

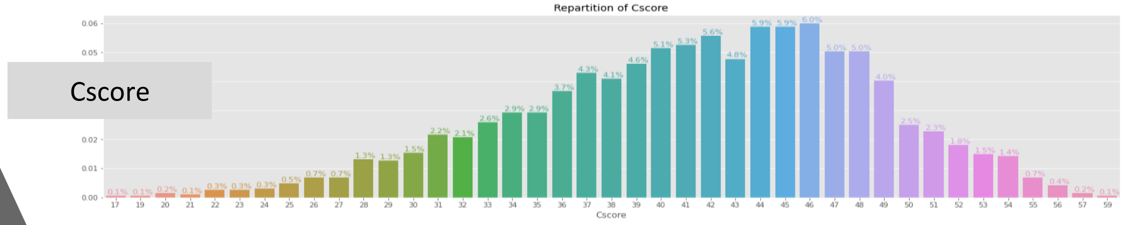
- We see that on all the score (N,E,O,C and A) the repartition of our population follow a gaussian distribution.

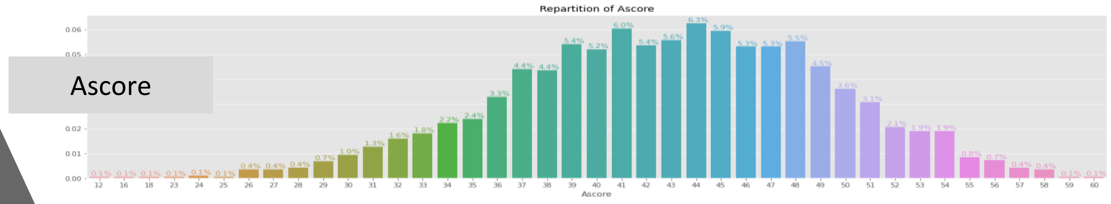
- Therefore, we can conclude that our population represents properly the entire population on a personal level as it is related to personality test









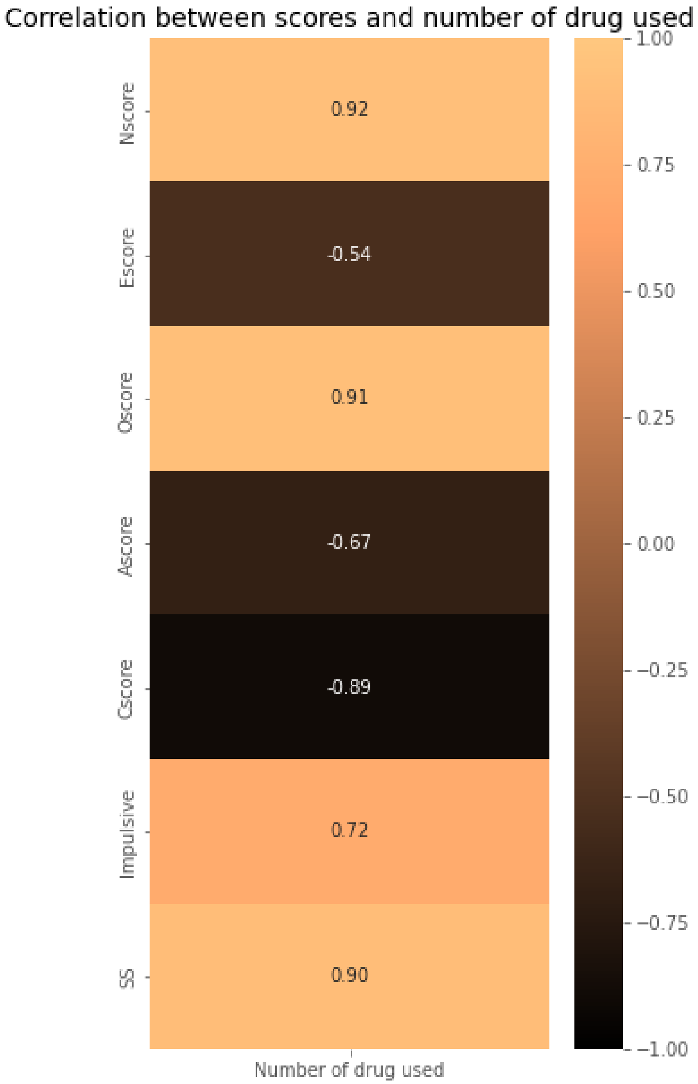


I – DATA VISUALISATION

- Here is a heatmap of the correlation between all scores of the 3 tests and number of drugs

- The authors of the survey showed that there is a relationship between risk of addiction to drugs and personality attributes, this could be confirmed with our heatmap, and we can observe a positive correlation for the N,O, and SS score and a negative correlation for the C score.

- This means for example, that the probability of having tested several drugs is increased by the more you are a sensitive person (linked to the N score).



I – DATA VISUALISATION

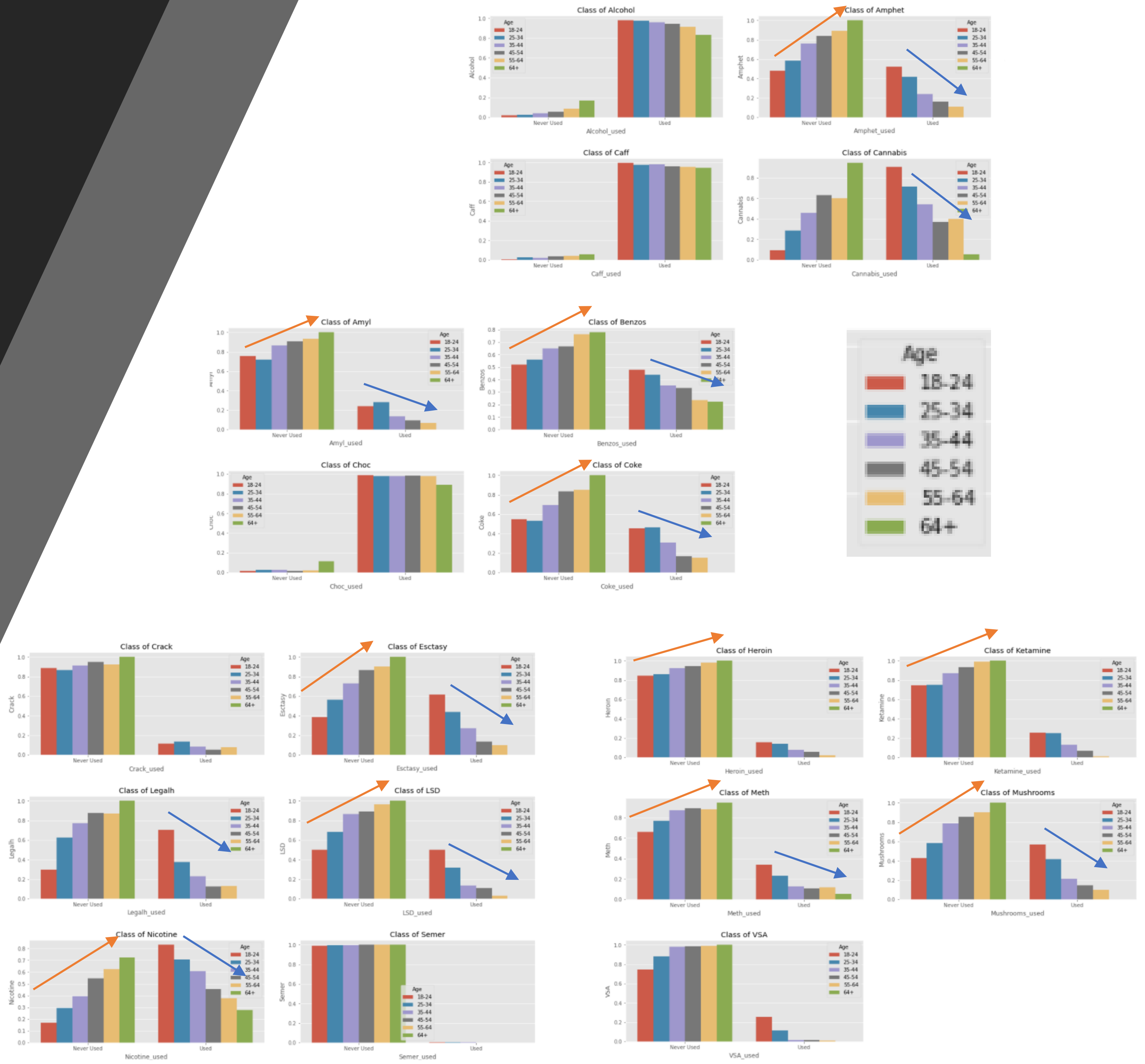
- Here is a heatmap of the correlation between all scores of the 3 tests and each drug independently

- This heatmap allows us to visualized which drug is impacted by which score.

I – DATA VISUALISATION

- Here are histograms showing the number of people using a particular drug categorized by groups of age, on the left we have the number of people whose never used the drug, and, on the left, we have the person who did

- As we can clearly see, underline with the red and bleu arrows, on almost every drug the amount of people using a drug is correlated with the age.



I – DATA VISUALISATION

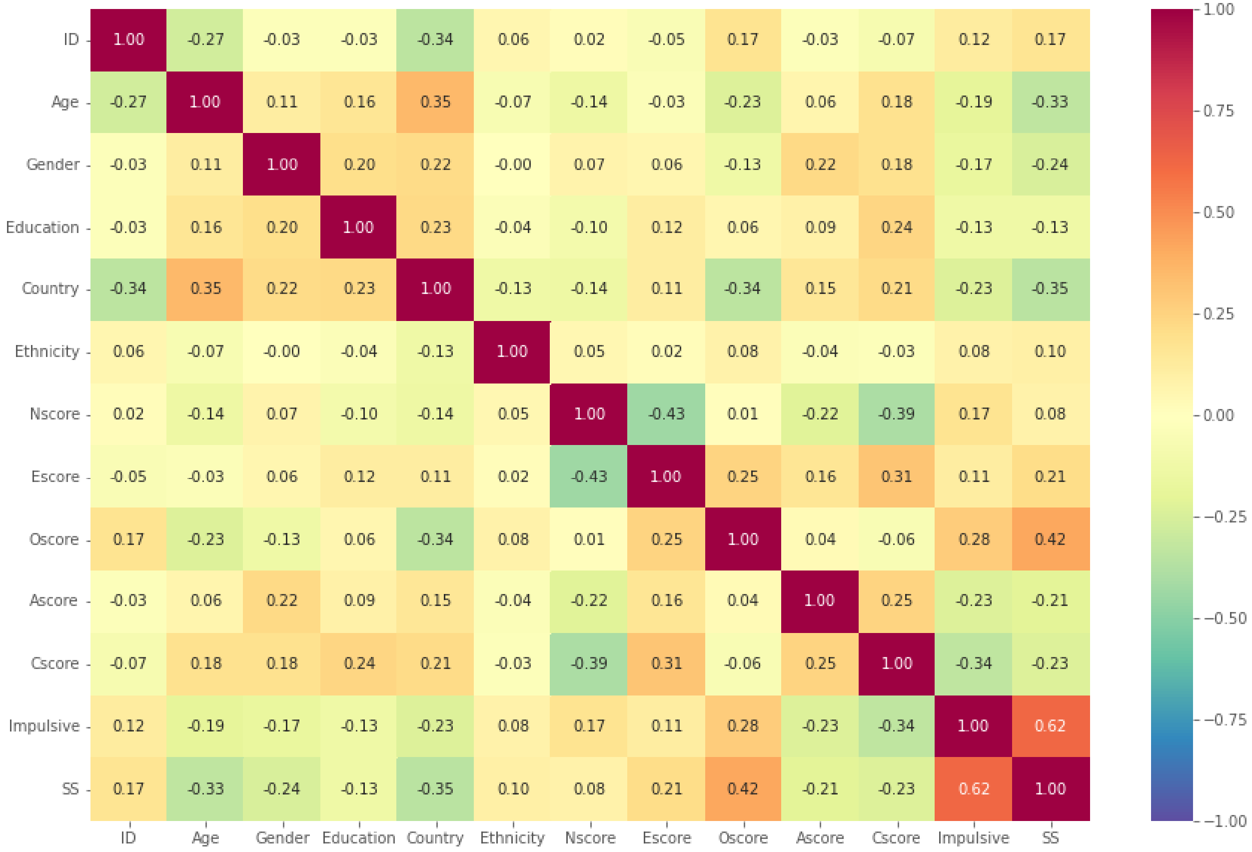
- Here are the repartition of our population in the different classes of drug consumption

- As we can observe there is no drugs where the repartition is homogenous, thus we will have to take this outbalanced problem in consideration during the processing part and the modelling one.

I – DATA VISUALISATION

- Here is a corrplot representation of all the features

- As we can see almost every features have weak correlation except for the Impulsive and SS feature. However, it is logic to have a correlation on these features because they are both representing the impulsivity of a person.



II – Preprocessing

1. Target preprocessing

- **Target selection**
  Drop Choc and Caff
- **Label encoding**

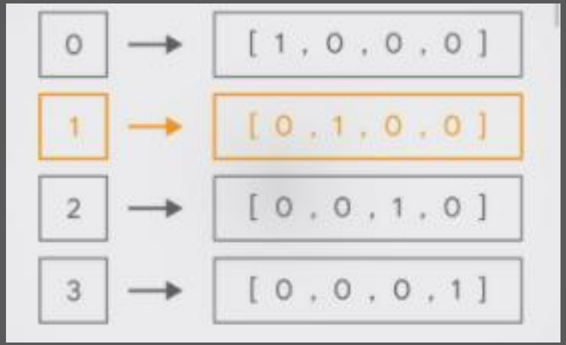
Class
CL0
CL1
CL2
CL3
CL4
CL5
CL6

Class
0
1
2
3
4
5
6

II – Preprocessing

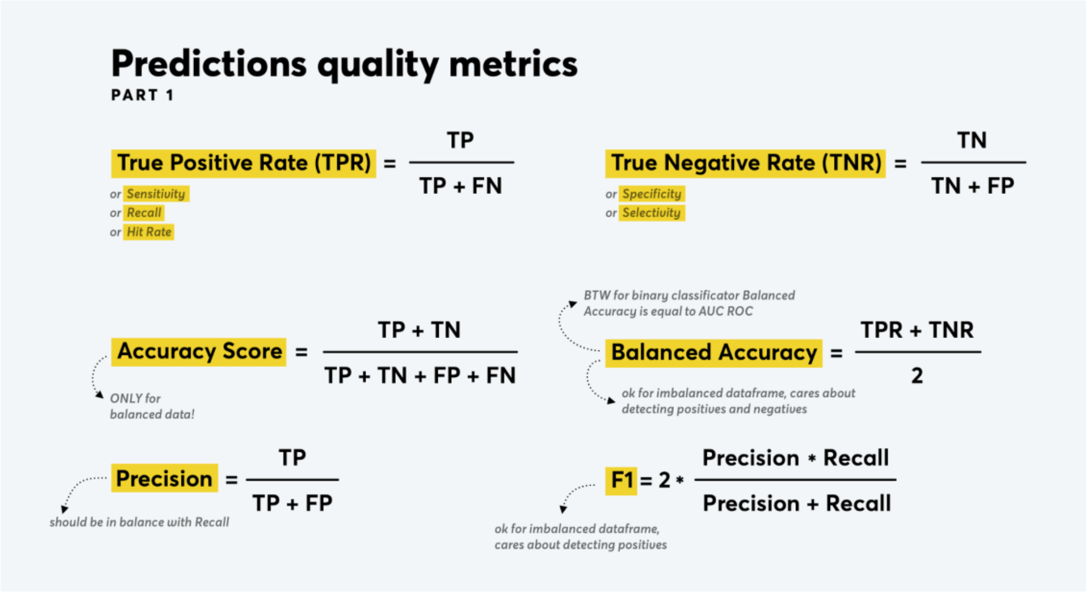
2. Features preprocessing

- **Feature selection**
  Drop feature ID
- **Feature encoding**
  - Encoding have already been performed on the original dataset
  - One hot encoding for country and ethnicity

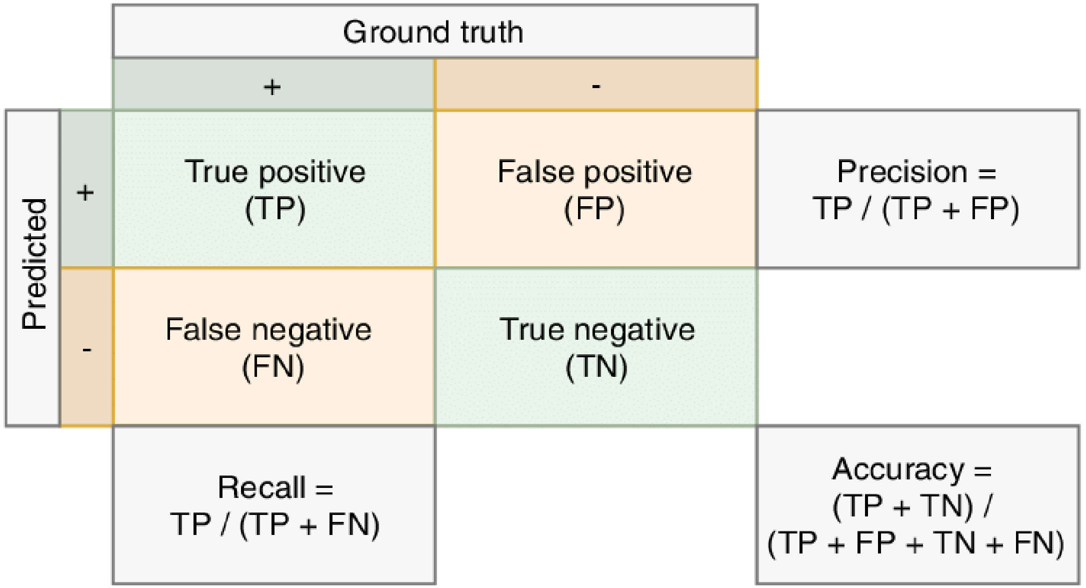


III- Modeling

Our Metrics

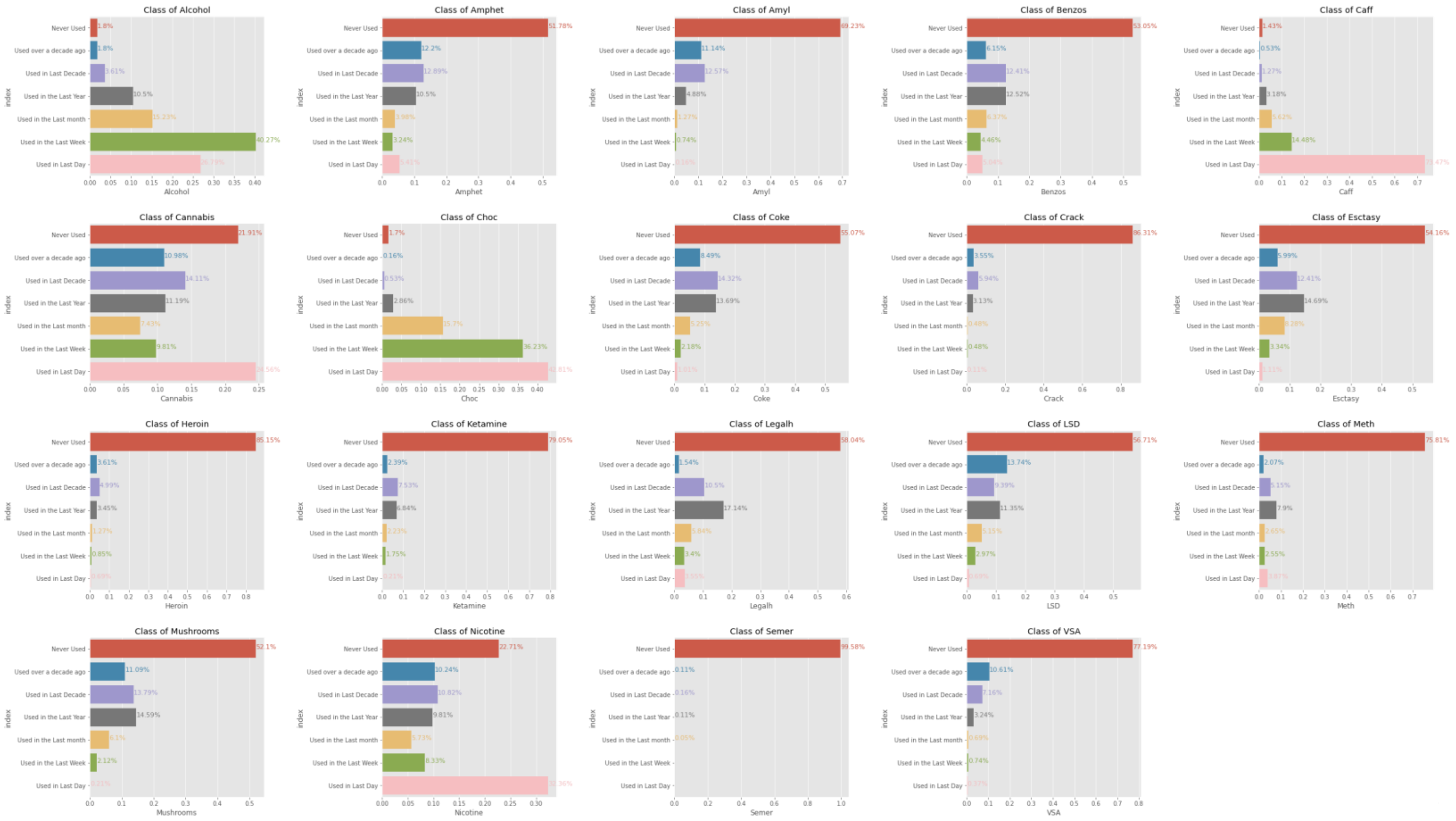


Confusion matrix



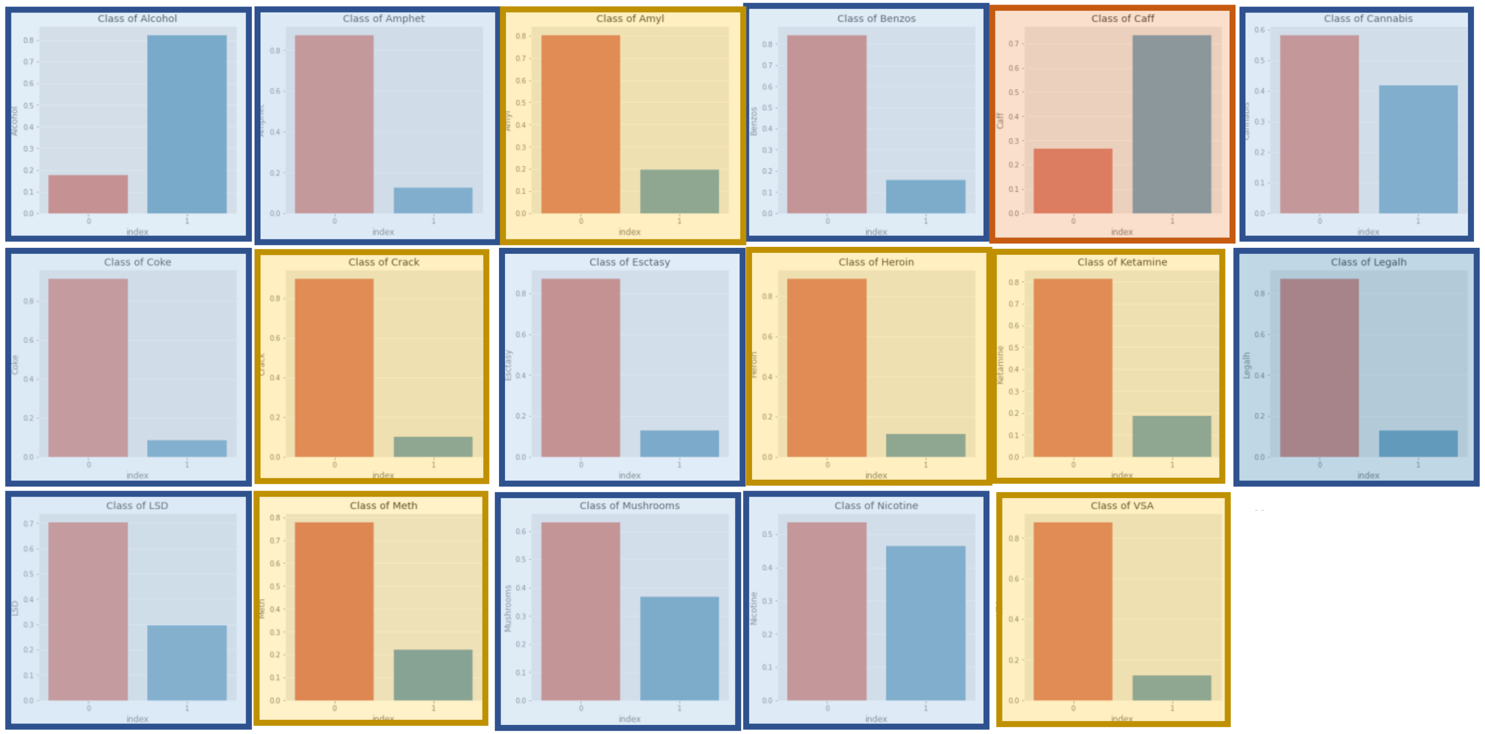
III- Modeling

Our Classes implementation: Initial classes



III- Modeling

Our Classes implementation: New classes



decade-based classification

Never Used (Or more than a decade ago)	Used

day-based classification

Not used yesterday	Used yesterday

month-based classification

Never used in the Last Month	Used in Last Month

III- Modeling

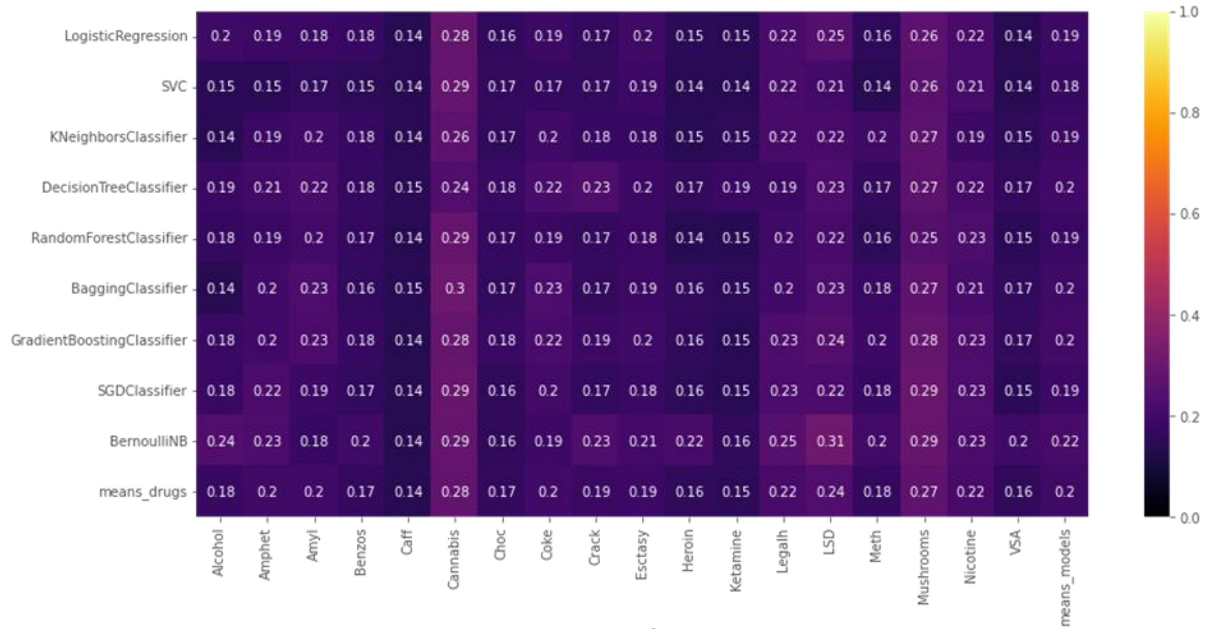
Our Approach

1. Find the best technic for imbalanced data
   - Base model
   - Weighting
   - Sampling (SMOTE)
2. Tuning hyperparameters with the best technic

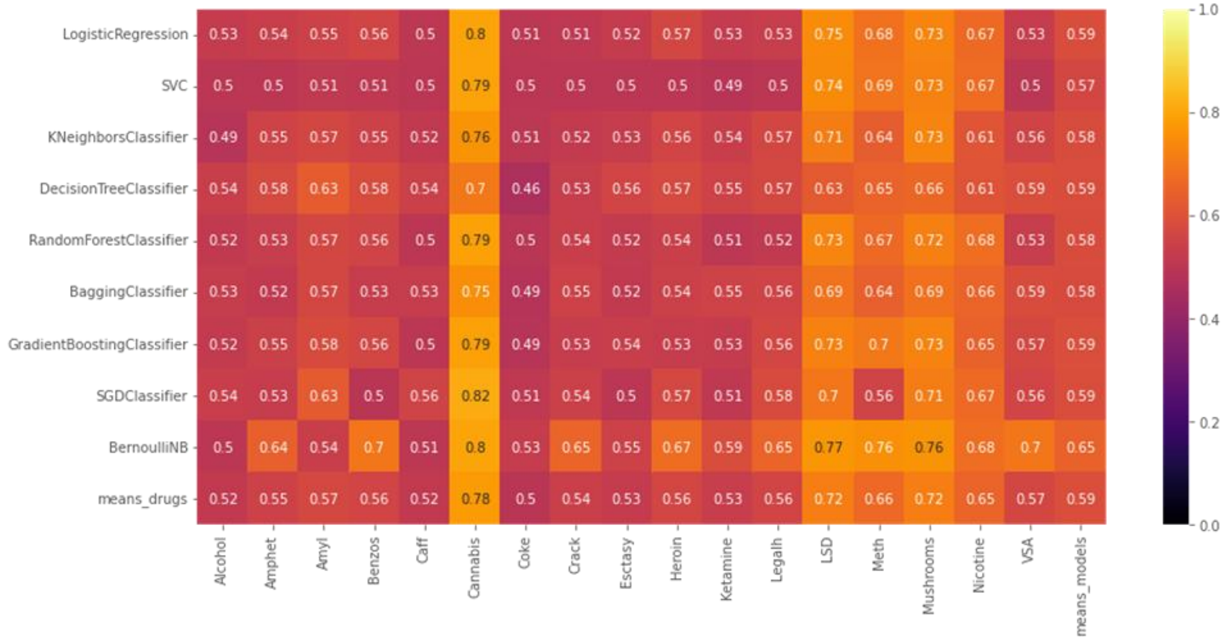
III- Modeling

Base Model

Initial classes

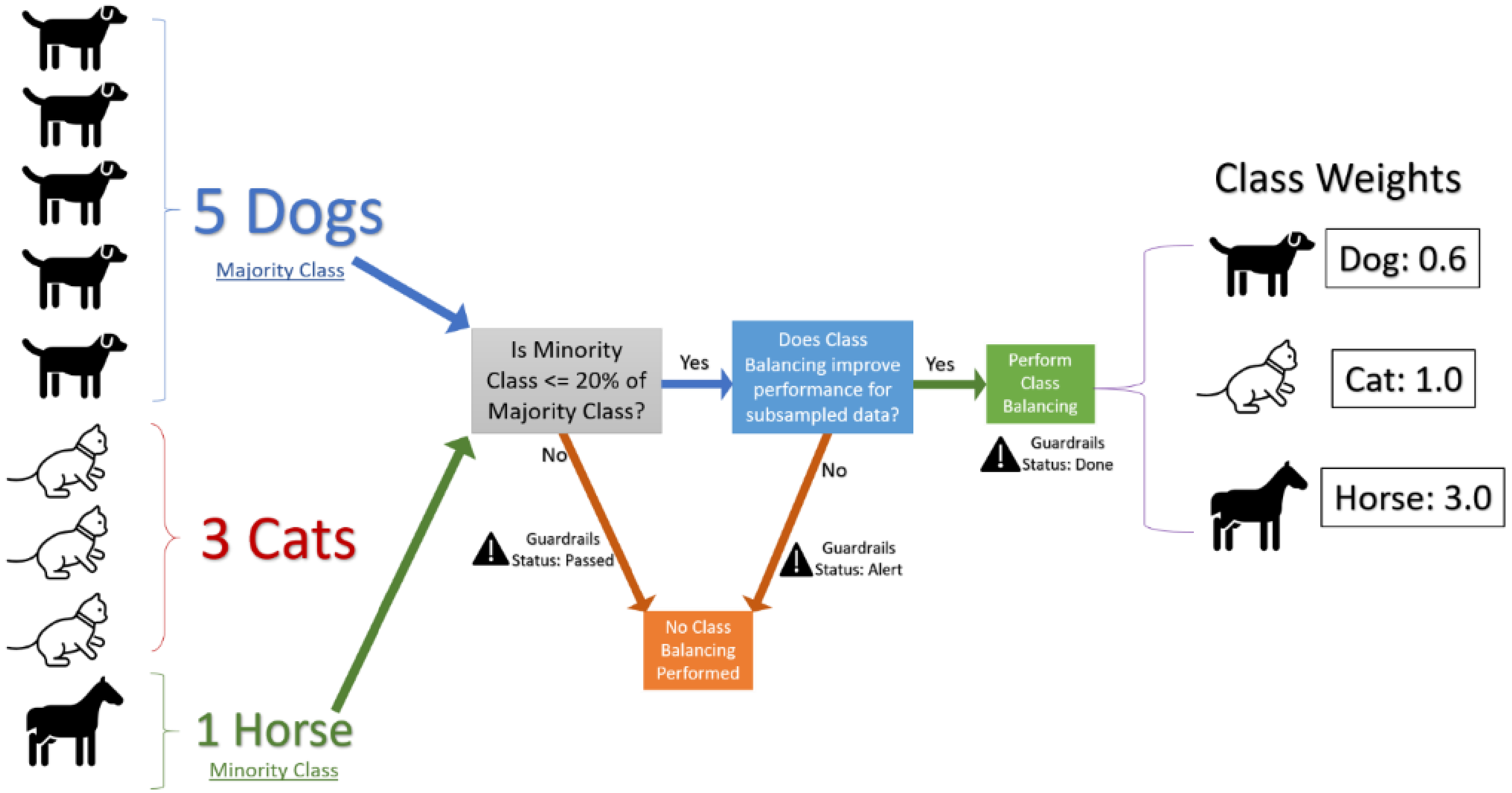


New classes



III- Modeling

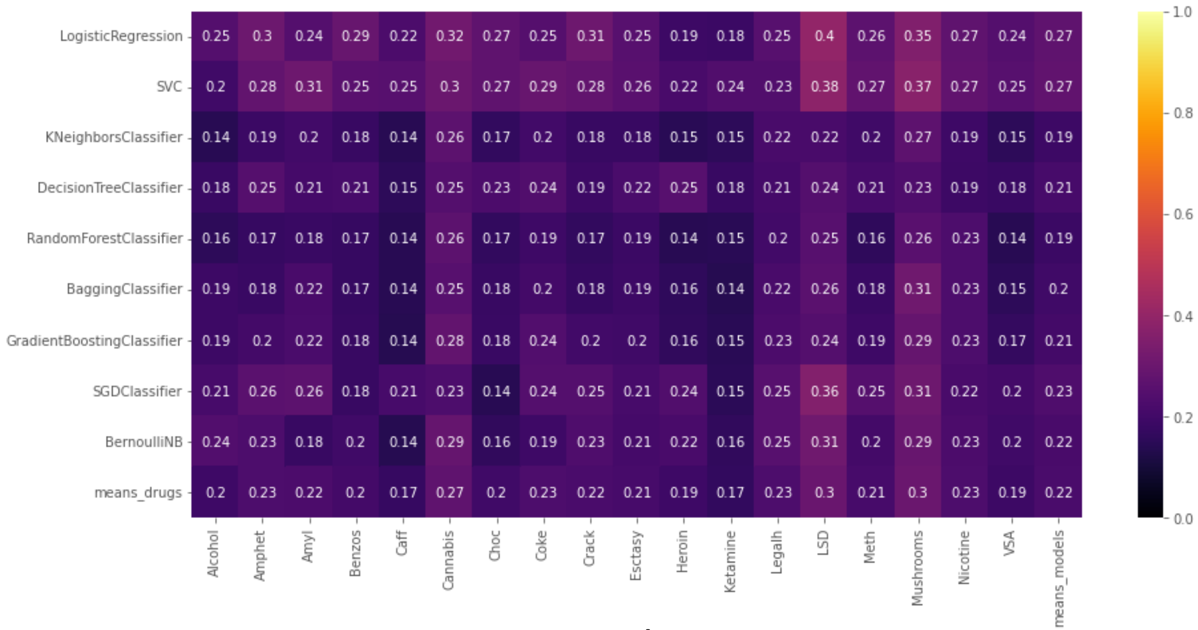
Weighting Explanation



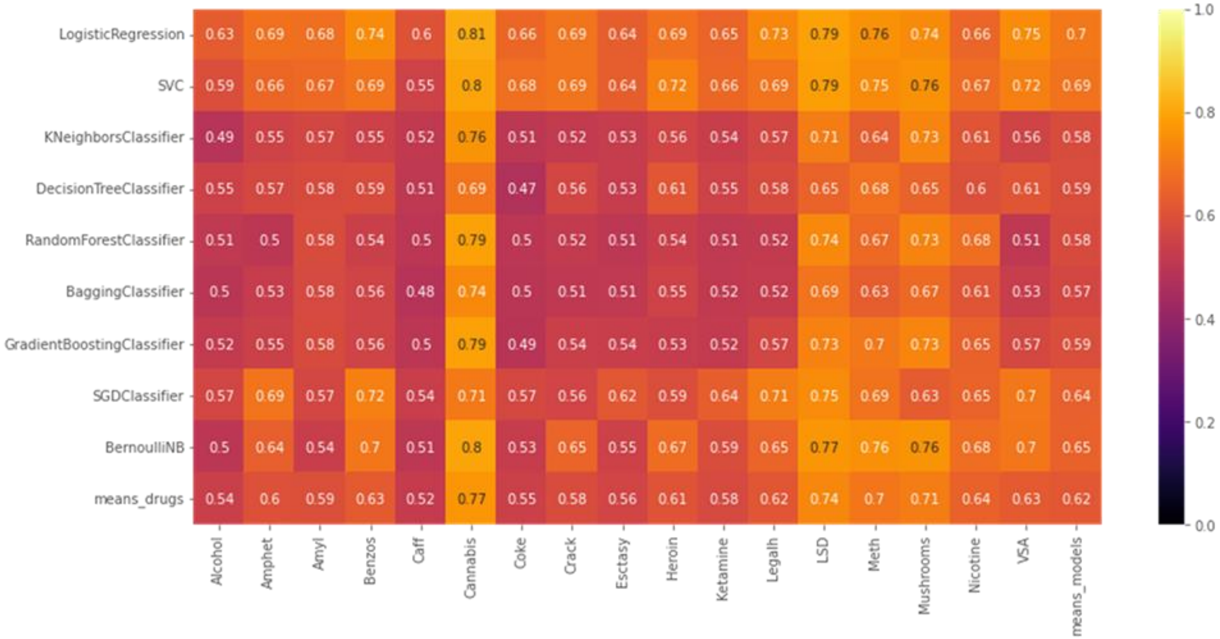
III- Modeling

Model using weighting

Initial classes

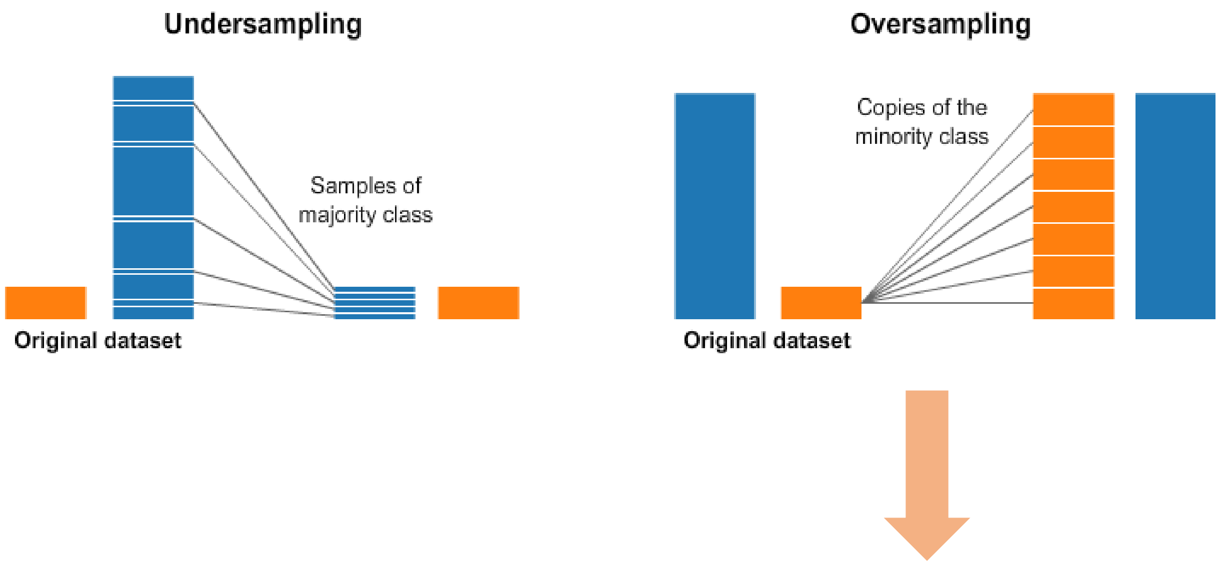


New classes

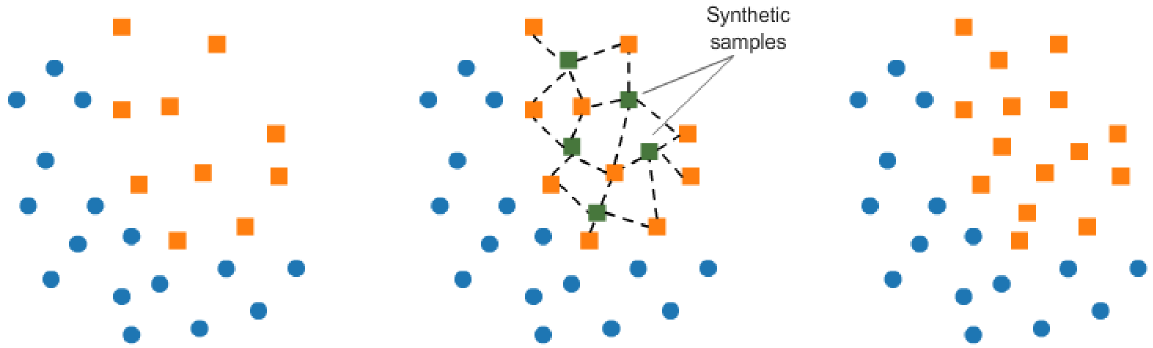


III- Modeling

Sampling Explanation



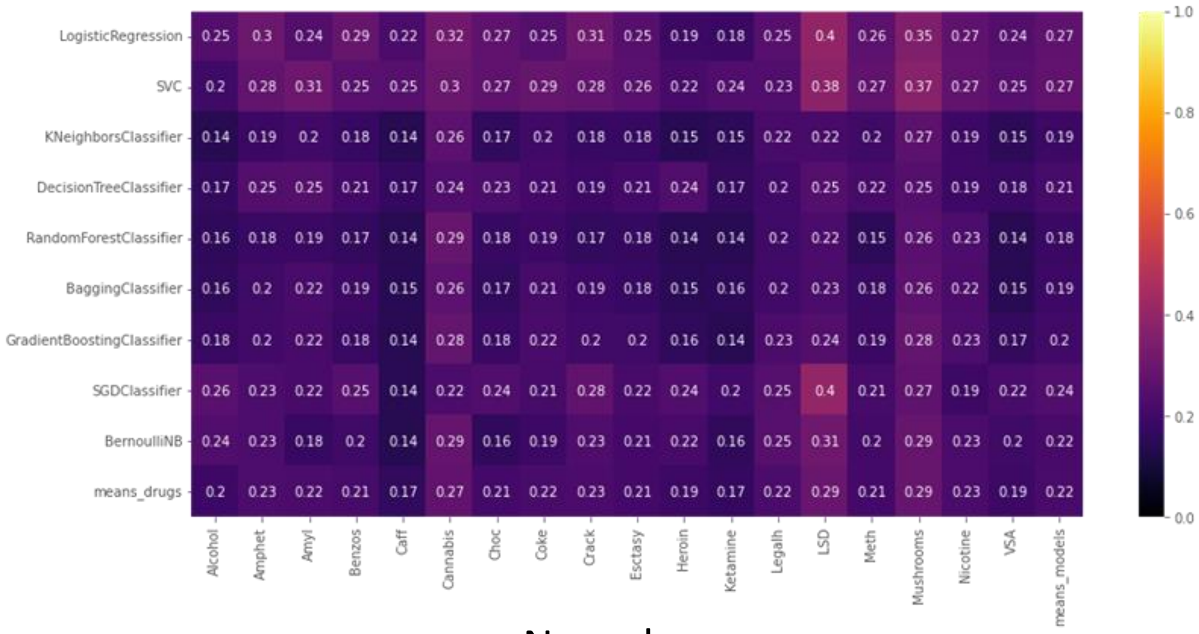
SMOTE (Synthetic Minority Oversampling Technique)



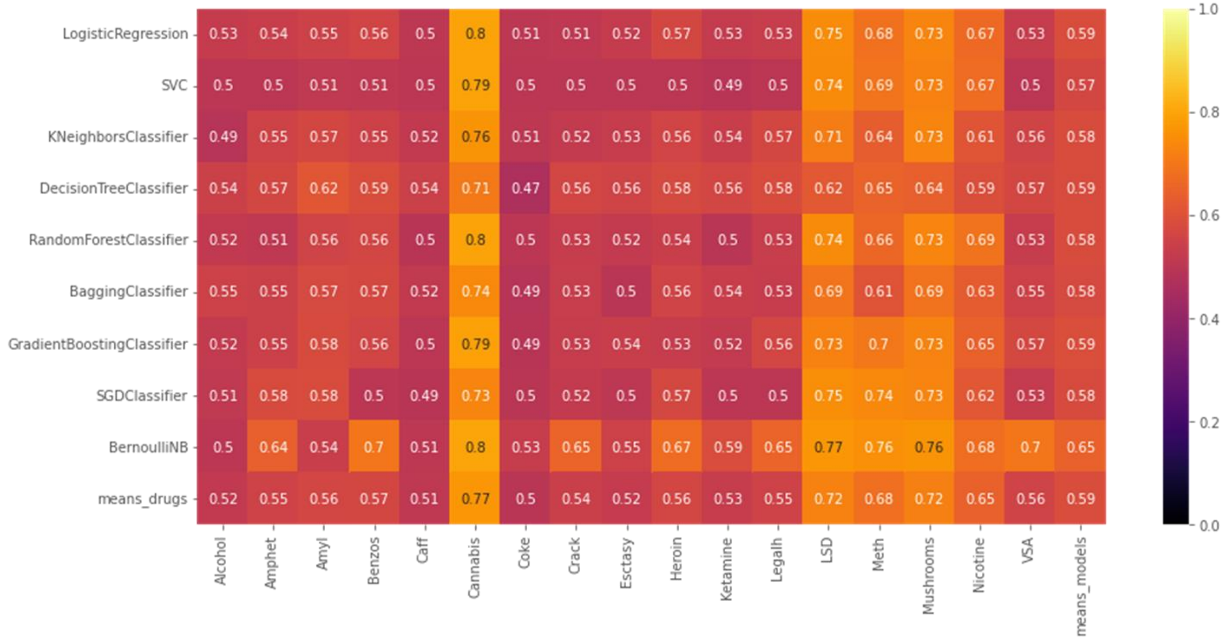
III- Modeling

Model using Sampling (SMOTE)

Initial classes



New classes



III- Modeling

Comparison

- Basic
- Weighting
- Sampling (SMOTE)

	Alcohol	Amphet	Amyl	Benzos	Caff	Cannabis	Coke	Crack	Esctasy
Basic	0.537057	0.641137	0.630848	0.696819	0.557714	0.816025	0.527858	0.654857	0.560399
Weight	0.629563	0.691905	0.680091	0.743058	0.602147	0.81445	0.680911	0.693601	0.640616
Sampling	0.548088	0.641137	0.6163	0.696819	0.543634	0.797769	0.527858	0.654857	0.557458
Basic_model_name	SGDClassifier	BernoulliNB	DecisionTreeClassifier	BernoulliNB	SGDClassifier	SGDClassifier	BernoulliNB	BernoulliNB	DecisionTreeClassifier
Weight_model_name	LogisticRegression	SGDClassifier	LogisticRegression	LogisticRegression	LogisticRegression	LogisticRegression	SVC	SVC	SVC
Sampling_model_name	BaggingClassifier	BernoulliNB	DecisionTreeClassifier	BernoulliNB	DecisionTreeClassifier	BernoulliNB	BernoulliNB	BernoulliNB	DecisionTreeClassifier
Best method	Weight_model_name	Weight_model_name	Weight_model_name	Weight_model_name	Weight_model_name	Basic_model_name	Weight_model_name	Weight_model_name	Weight_model_name
Best model	LogisticRegression	SGDClassifier	LogisticRegression	LogisticRegression	LogisticRegression	SGDClassifier	SVC	SVC	SVC

	Heroin	Ketamine	Legalh	LSD	Meth	Mushrooms	Nicotine	VSA
Basic	0.672854	0.589731	0.653969	0.772223	0.755291	0.762485	0.679324	0.697908
Weight	0.724158	0.663761	0.734438	0.792161	0.760893	0.762485	0.679324	0.747909
Sampling	0.672854	0.589731	0.653969	0.772223	0.755291	0.762485	0.689977	0.697908
Basic_model_name	BernoulliNB	BernoulliNB	BernoulliNB	BernoulliNB	BernoulliNB	BernoulliNB	BernoulliNB	BernoulliNB
Weight_model_name	SVC	SVC	LogisticRegression	LogisticRegression	LogisticRegression	BernoulliNB	BernoulliNB	LogisticRegression
Sampling_model_name	BernoulliNB	BernoulliNB	BernoulliNB	BernoulliNB	BernoulliNB	BernoulliNB	RandomForestClassifier	BernoulliNB
Best method	Weight_model_name	Weight_model_name	Weight_model_name	Weight_model_name	Weight_model_name	Basic_model_name	Sampling_model_name	Weight_model_name
Best model	SVC	SVC	LogisticRegression	LogisticRegression	LogisticRegression	BernoulliNB	RandomForestClassifier	LogisticRegression

III- Modeling

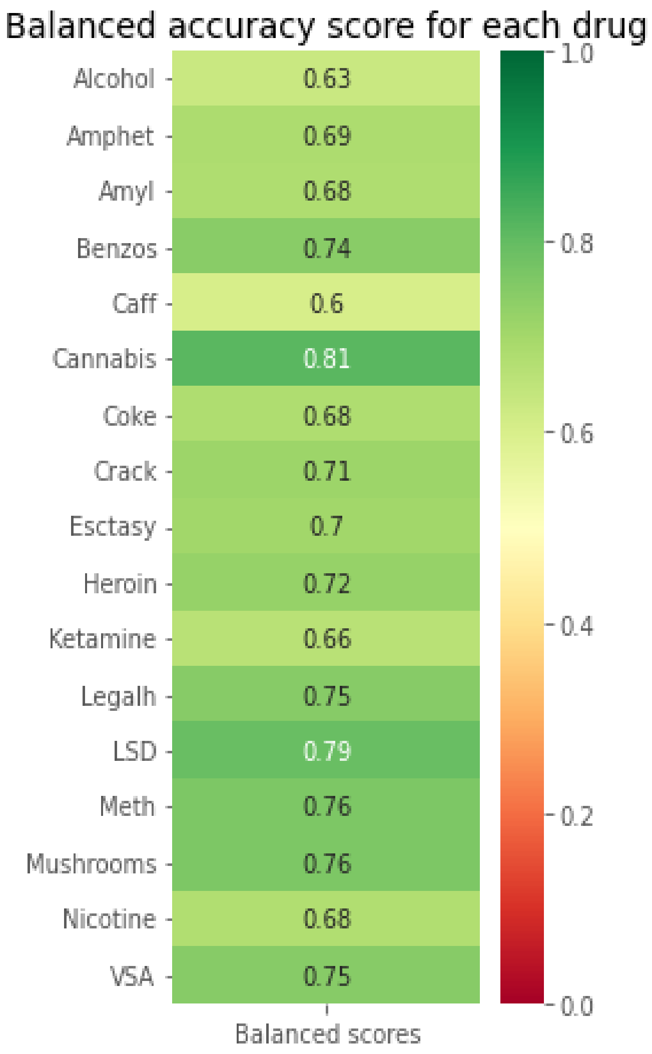
Tuning Hyperparameters Result

IV- Final Models

We have achieved a mean of 71.24% of balanced accuracy.

Logistic Regression, Support Vector Machine and Bernoulli NB are the best models for the drugs.

Drug	Model
Alcohol	Logistic Regression
Amphet	Logistic Regression
Amyl	Logistic Regression
Benzos	Logistic Regression
Caff	Logistic Regression
Cannabis	Logistic Regression
Coke	SVC
Crack	Logistic Regression
Esctasy	SVC
Heroin	SVC
Ketamine	SVC
Legalh	SVC
LSD	SVC
Meth	Logistic Regression
Mushrooms	BernoulliNB
Nicotine	BernoulliNB
VSA	Logistic Regression



Thank you for you attention and for a careful reading