SORBONNE UNIVERSITÉ

Drug Consumption Dataset

MACHINE LEARNING FOR MEDICINE

Romain CHOR & Mohamed RUINE

Master 2 in Statistics — March 31st, 2020

Outline

I. Introduction

II. Standard Machine Learning methods

III. Optimal Solution

IV. Conclusion

I. Introduction

Dataset Description

- **UCI Dataset :** Drug Consumption (quantified)
- **Relevant Paper :** Fehrman E., Muhammad A.K., Mirkes E.M., Egan V., Gorban A.N. (2017) "The Five Factor Model of Personality and Evaluation of Drug Consumption Risk"
- **Abstract :** Classify type of drug consumer by personality data
- **Database Origin :** Online survey held during 2011
- **Database Characteristics :**

Associated Task :	Classification	**Number of Samples :**	1,885
Attribute Types :	Real ($\mathbb{R}$)	**Number of Features :**	32

I. Introduction

Dataset Description

- For each of the 1,885 respondents, 12 attributes are known :
  - 5 (quantified) categorical features : Age (band), Gender, Level of Education, Country of Residence, Ethnicity
  - 7 (continuous) numerical features, corresponding to personality traits and temperament :
    - Neuroticism (**N**), revealing a sensitive/nervous vs. secure/confident profile
    - Extraversion (**E**), revealing an outgoing/energetic vs. solitary/reserved profile
    - Openness (**O**), revealing an inventive/curious vs. consistent/cautious profile
    - Agreeableness (**A**), revealing a friendly/compassionate vs. challenging/detached profile
    - Conscientiousness (**C**), revealing an efficient/organized vs. easy-going/careless profile

    - Impulsivity (**Imp**)
    - Sensation-Seeking (**SS**)

I. Introduction

Problem Description

- In addition, participants were questioned concerning their use of 18 legal and illegal drugs : Alcohol, Amphetamines, Amyl nitrite, Benzodiazepine, Cannabis, Chocolate, Cocaine, Caffeine, Crack, Ecstasy, Heroin, Ketamine, Legal highs, LSD, Methadone, Mushrooms, Nicotine and Volatile substance abuse. Finally, one fictitious drug (Semeron) which was introduced to identify over-claimers.

- For each drug, they have to select one of these 7 answers :
  - never used the drug, or used it over a decade ago,
  - or in the last decade, year, month, week, or day.

I. Introduction

Problem Description

- Therefore, one can see the database as 18 classification problems !
- Problem which can be solved as a :
  - 7-class classification for each drug separately, or
  - bi-class classification by union of part of classes into one new class.

Value	7-Class Description	Bi-Class Description
CL0	Never Used	Not User (0/False)
CL1	Used over a Decade Ago	
CL2	Used in Last Decade	(Drug) User (1/True)
CL3	Used in Last Year	
CL4	Used in Last Month	
CL5	Used in Last Week	
CL6	Used in Last Day	

I. Introduction

Dataset Exploration

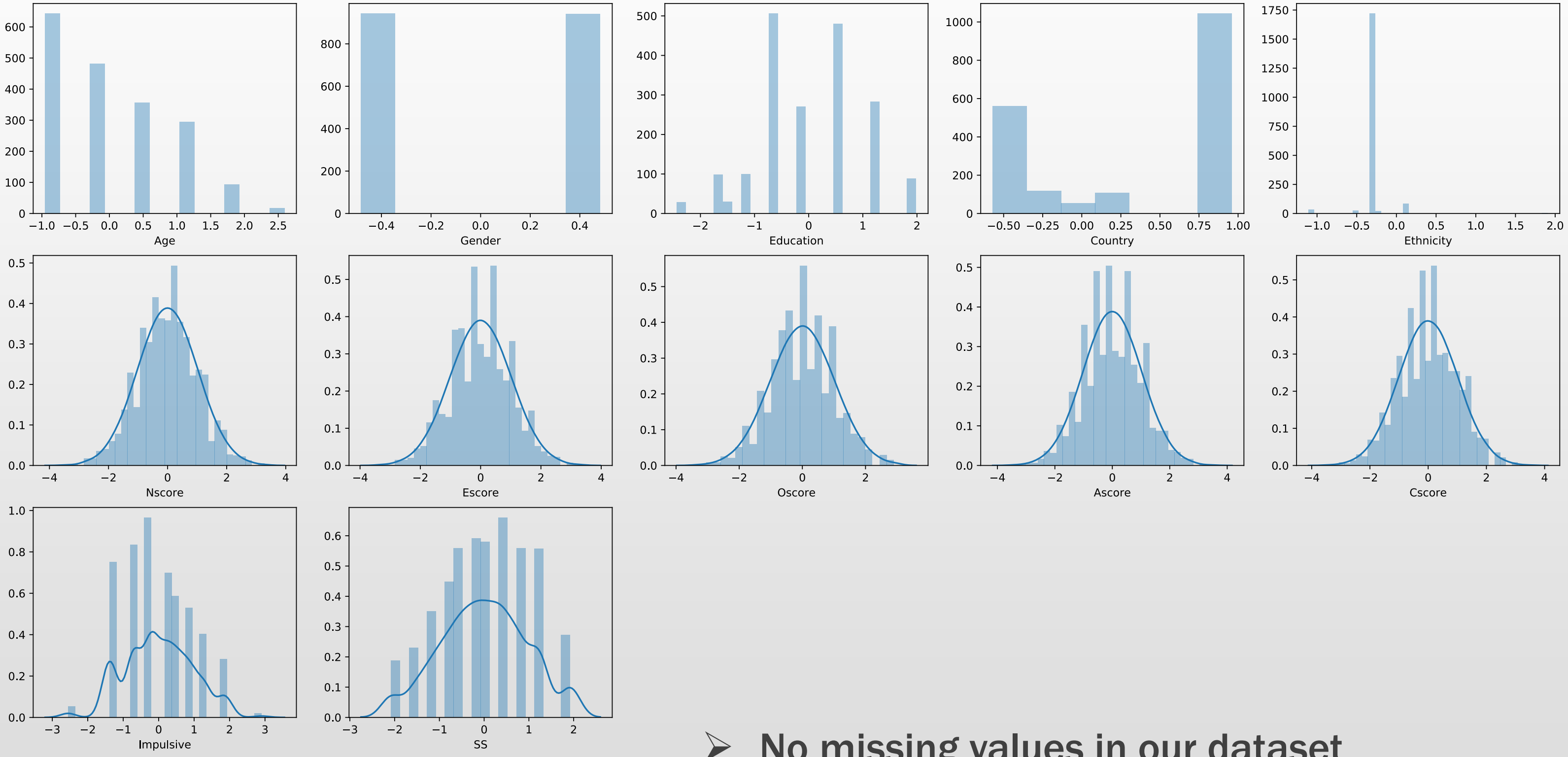

Fig 1. Distribution plots for the 12 attributes

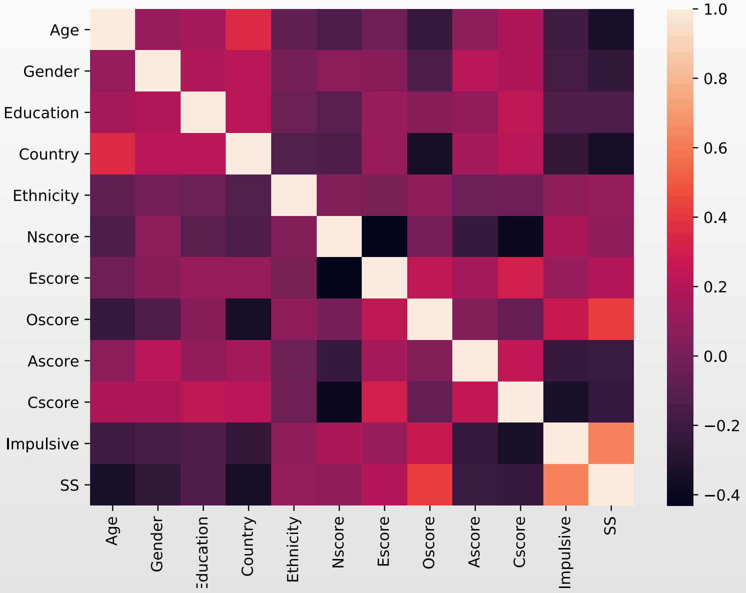

Fig 2. Heat Map for the 12 attributes

- ➢ No missing values in our dataset
- ➢ Personality scores already have standard normal distributions
- ➢ No remarkable correlation between features except for *Impulsive* and *SS* scores

I. Introduction

Dataset Exploration

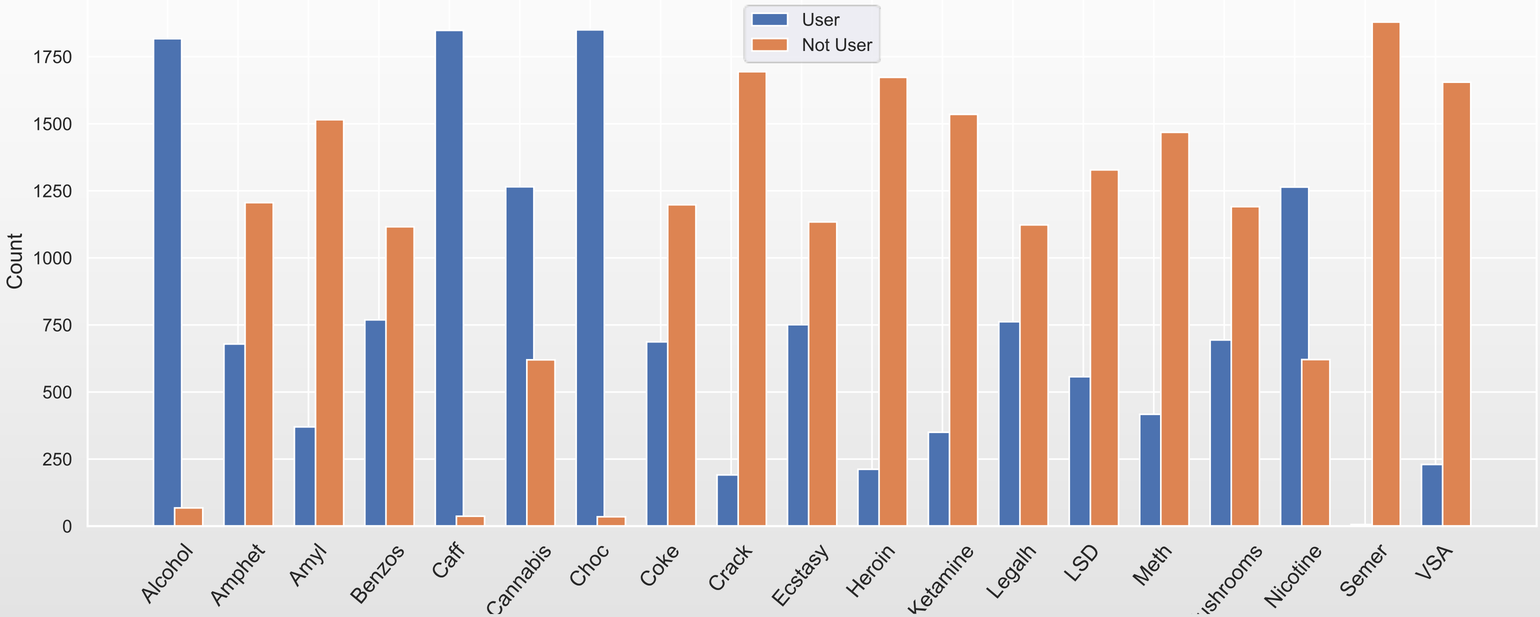

Fig 3. Distribution plots for the 18 drugs

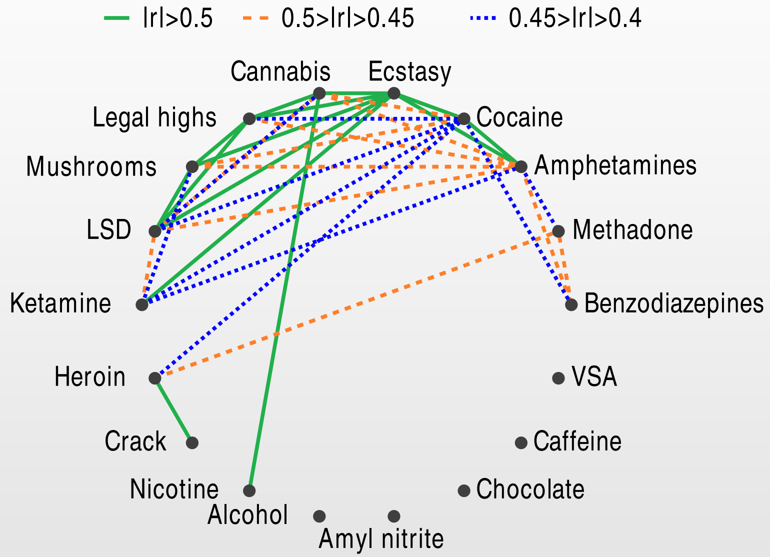

Fig 4. Correlation Links for the 18 drugs

- ➢ Some drugs have particularly unbalanced distributions such as *Heroin* or *Crack*. The remaining ones like *Alcohol* are naturally obvious. Treating them will require special techniques such as resampling.
- ➢ One can observe that some drugs are quite correlated to each other. This means that a person consuming a drug A is likely to consume a drug B as well. A study using only drugs as features can be conducted but this is out of our subject of interest.

I. Introduction

Dataset Exploration

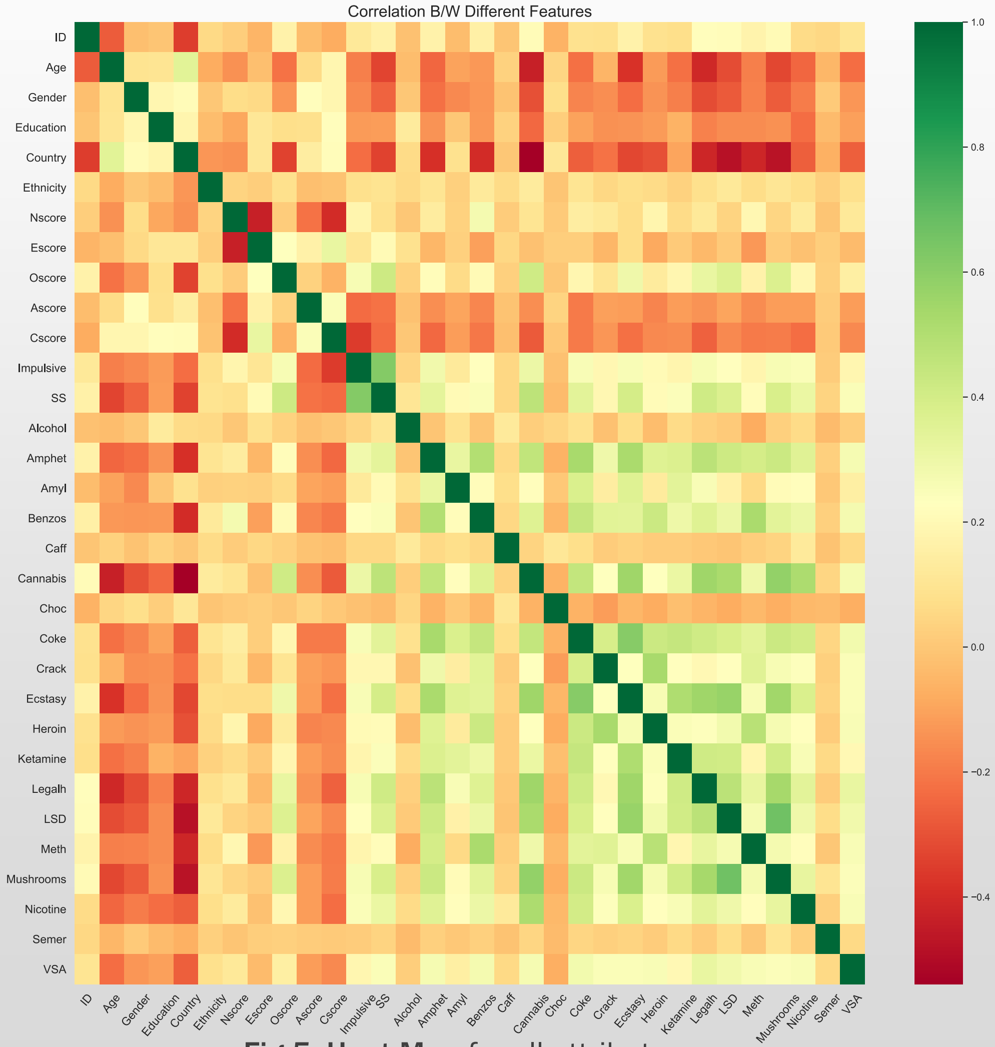

Fig 5. **Heat Map** for all attributes

- One way to make predictive modelling is to use all available features. This way we can exploit correlations between the different drugs consumption.

- However let us first try to predict without these information.

- As an example, we first focus on the LSD drug. Indeed, it is one of the few relatively well-balanced class, that is also **not** highly correlated with one the base attributes.

I. Introduction

Dataset Exploration

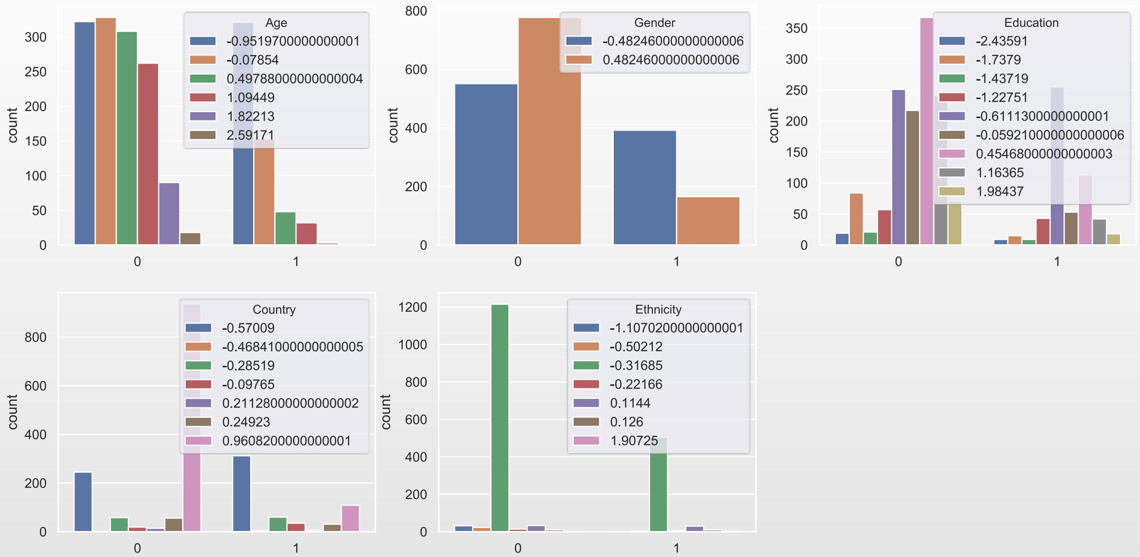

Fig 6. Count plots for the categorical attributes

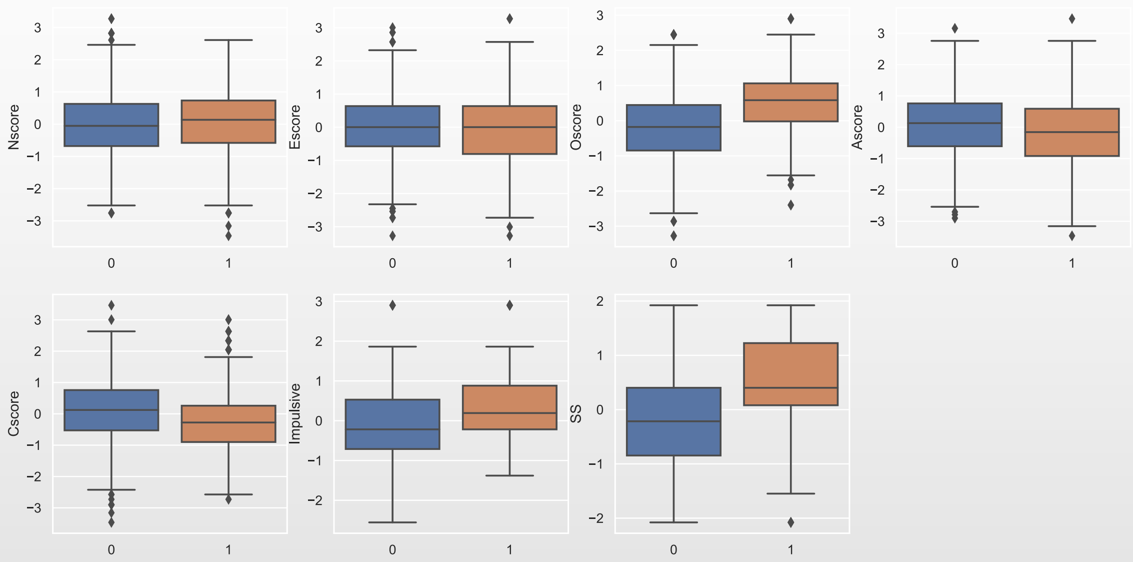

Fig 7. Box plots for the numerical attributes

- ➢ **Count plots show that distributions for the categorical features are different depending on the target.**
- ➢ **However, note that we do not observe different distributions for the continuous features.**

II. Standard Machine Learning methods

Basic models

- For our LSD classification problem, let us run some state-of-the-art ML methods, so we can have an idea on the quality of our dataset.
  - **Linear Classifiers :** Logistic Regression (LR), Support Vector Classification (SVC)
  - **Nearest Neighbors :** K-NN
  - **Ensemble Methods :** Random Forests (RF), Extra-Trees (ExT), Gradient Boosting (GB), XGBoost, LightGBM

	LR	SVM	KNN	RF	ExT	GB	XGB	LGBM
CV-Run0	0.877671	0.861951	0.819354	0.839716	0.860052	0.864747	0.814528	0.828559
CV-Run1	0.851559	0.846284	0.794271	0.841351	0.845202	0.845440	0.813103	0.829984
CV-Run2	0.844130	0.850314	0.826336	0.840304	0.848742	0.835299	0.824790	0.825314
CV-Run3	0.865169	0.857378	0.836840	0.856053	0.853376	0.848156	0.828705	0.824359
CV-Run4	0.862042	0.848315	0.822636	0.859869	0.846937	0.865593	0.830772	0.835542
Mean score	0.860114	0.852848	0.819888	0.847459	0.850862	0.851847	0.822380	0.828752
Std	0.011543	0.005890	0.014093	0.008675	0.005342	0.011695	0.007266	0.003969
Training time	1.434549	0.088302	0.072496	0.556560	0.459513	0.389365	1.574409	0.249584

II. Standard Machine Learning methods

Basic models

- Implementation of a 5-Fold Cross Validation.
- Use of the ROC AUC metric to evaluate our models. Indeed, for classification of imbalanced targets, accuracy score performs poorly and does not help distinguish model performances.
- A default Logistic Regression maximized our metric with a score of ~86%.
- Let us try to add feature engineering techniques to improve the score.

II. Standard Machine Learning methods

Feature Engineering

- **Standardization :** Shift feature values towards the center and standardize them
  - Removing the mean and scaling to unit variance $\frac{X-\mu}{\sigma^2}$ to approach a normal distribution (as the personality scores)
    - ➢ After re-running the same state-of-the-art methods, standardization seems to degrade performances.

- **Feature Imputation :** Add some features to pass on more relevant information to the model
  - Number of drugs consumed
  - Mean and Variance of the personality scores
    - ➢ After re-running the same methods, imputation seems to upgrade performances (~86.1%).

II. Standard Machine Learning methods

TP/TME Methods

- **Decision Trees** (TP1) :
  - The tree obtained has of course an accuracy of 1.0, but we can see in the graph result that the decision tree is highly complex, which obviously indicate overfitting.

- **Clustering** (TP2) :
  - Even if we choose clustering methods reacting to many different use cases, we still get very poor results, which could indicate that our dataset has no relevant shape regarding to the features' space.

- **PCA** (TP6) :
  - Since the number of features at hand is not very high for our drug dataset, PCA has only a harmful effect. Thus, applying no dimensionality reduction technique remains to be the best solution.

II. Standard Machine Learning methods

Neural Nets

- Artificial Neural Networks (ANN) have the reputation to be efficient with binary classification, especially when the data is standardized.

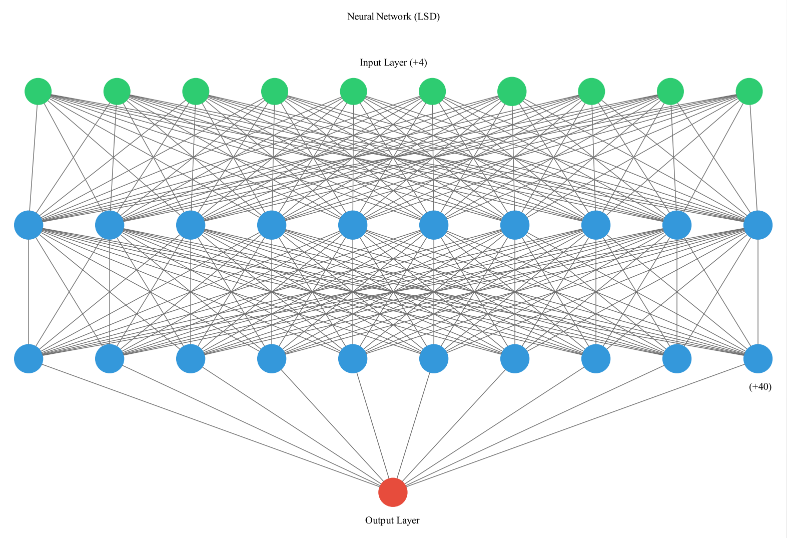

Fig 8. ANN Architecture

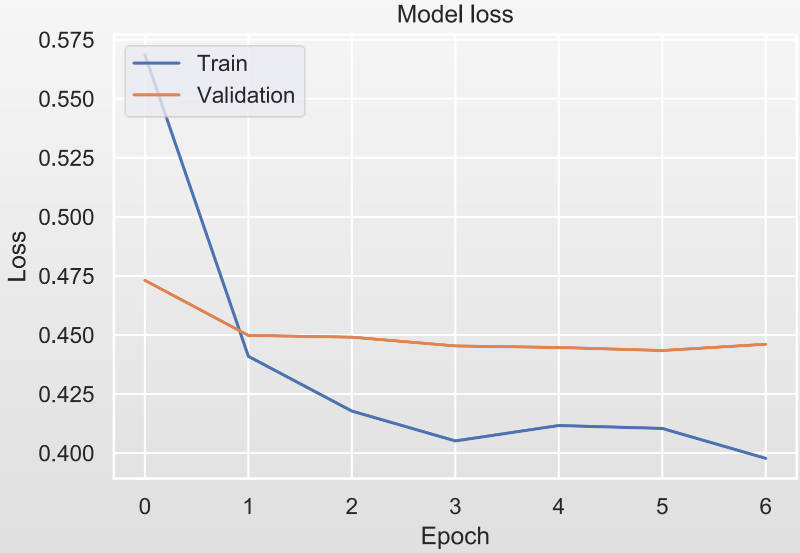

Fig 9. Losses plots for the ANN

- ➢ Note that neural networks in our case are very prone to overfit. We can indeed observe that when the NN has too many layers or too many neurons per layer, validation loss increases while training loss decreases. The dropout layer helps prevent overfitting.
- ➢ This explains the difficulty to obtain better performances than classical ML models.

III. Optimal Solution

Aggregation Methods

- **Blending** consists in using multiple models to predict a target, then getting the mean of those predictions.
  - Combination used : Logistic Regression, Random Forest, Gradient Boosting, XGBoost
  - With this combination of models, blending performs better than all single models

	LR	KNN	RF	ExT	GB	XGB	LGBM	Blending	Stacking	NN
Score	0.816331	0.777552	0.811895	0.814519	0.824020	0.778195	0.797975	0.821141	0.782429	0.789779
Training time	0.009972	0.005009	0.231758	0.170028	0.242192	0.168419	0.110278	0.721152	3.758186	1.514289

- Let us see what happens when trying to predict our LSD target using all other drugs :

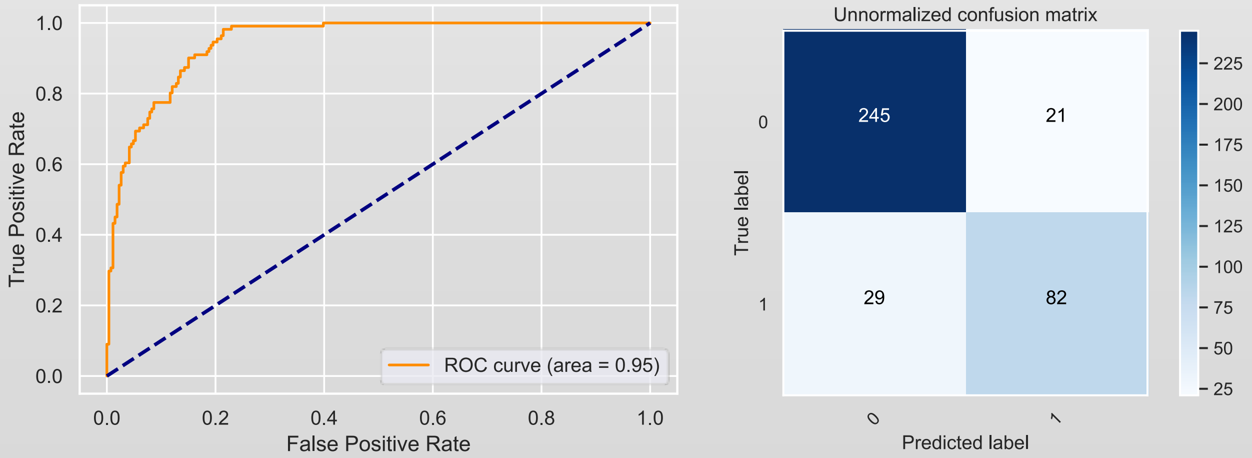

Fig 10. ROC Curve regarding TPR and FPR

- ➢ We gain up to **12%** for our ROC AUC score which is huge !
- ➢ Overall performances are great so conducting prediction using all drugs features yields no challenge.

IV. Conclusion

Summary

- What we achieved :
  - Established a benchmark with several classical and state-of-the-art ML models
  - Tried to improve performances with Neural Networks and Aggregation techniques
  - Models aggregation techniques like Blending worked well
  - Proved the relation between the personality scores and drug consumption

- Encountered Difficulties :
  - What type of classification problem should one choose ?
    - Multi-class, bi-class, regression ?
    - Which drug(s) ? How many of them to keep or to withdraw ?
  - Neural Networks were prone to overfit
  - The reference was very long to read, and sometimes hard to understand

IV. Conclusion

Discussion

- About the reference :
  - Very simple methods were used with good results
    - As such, in our opinion, their problem was oversimplified (binary classification, drug clusters, …)
    - Moreover, they ran thousands of different combinations to achieve such metrics

- Paths for further analysis :
  - Hyperparameters tuning, especially for NN
  - Use other targets, especially unbalanced targets with corresponding solutions
  - Return to the original multi-class classification problem