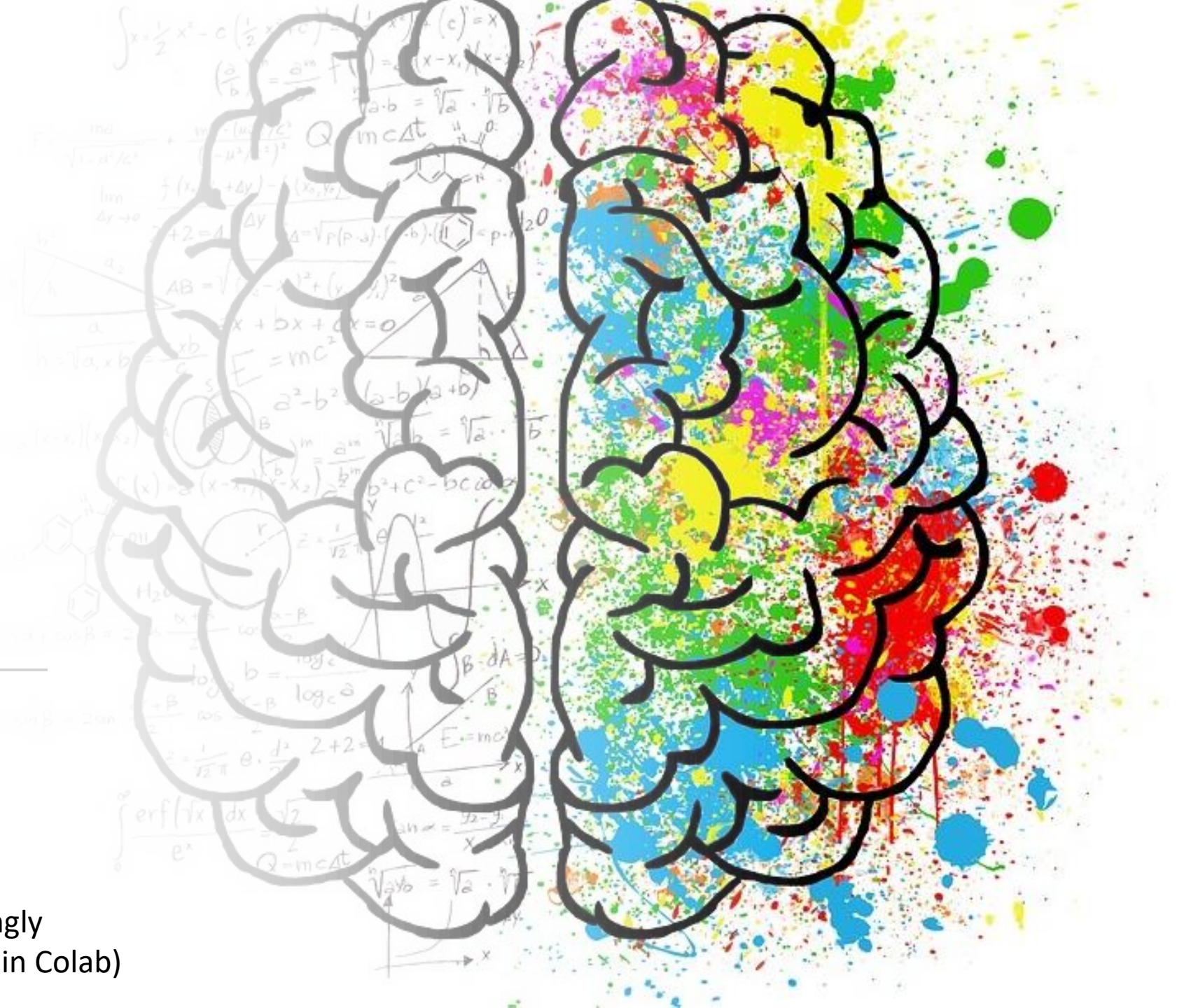
Drug consumption (quantified) Data Set

Classify type of drug consumer by personality data

ADJARIAN Stéphan – Notebook (I strongly recommend to take a look at the code in Colab)



Ins and Outs of the problem

Database contains records for 1885 respondents. For each respondent 12 attributes are known: Personality measurements which include NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity. All input attributes are originally categorical and are quantified. After quantification values of all input features can be considered as real valued. 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 and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug they have to select one of the answers: never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day.

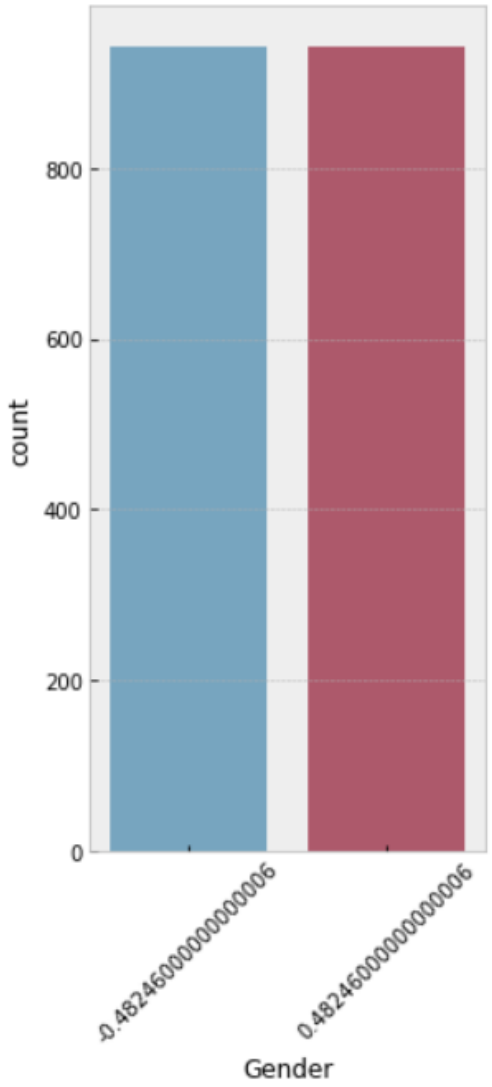
Database contains 18 classification problems. Each of independent label variables contains seven classes. We transformed the 7-classification problem to **binary classification** by union of part of classes into one new class. "Never Used", "Used over a Decade Ago" and "Used in Last Decade" form class "Non-user" and all other classes form class "User".

Ins and Outs of the problem

We want to predict if the person is likely to use a certain drug in function of personality measurements. We want to assess how correlated the personality is to drug consumption. To do so, we tried various famous machine learning algorithm and classification methods (multinomial logistic regression, decision tree, random forest, boosting, k-nearest neighbors) and the most effective classifier was selected for each drug. We used Holdout Method to validate the accuracy obtained and we performed a grid search algorithm to select the best hyperparameters.

I was strongly inspired by work of the scientists Elaine Fehrman, Evgeny M. Mirkes, Awaz K. Muhammad, Vincent Egan and Alexander N. Gorban who conducted the research for *The Five Factor Model of personality and evaluation of drug consumption risk*.

My thoughts are that drug consumption is highly defined by our character. On the graph on the right we observe that quota method with respect to gender has been used to create this dataset *(-0.48 stands for male and 0.48 stands for female).* We will see in which extent it is possible to predict if a person is a drug user (used less than a year ago) and this for each drug.



Results from *The Five Factor Model of personality and evaluation of drug consumption risk*

Table 18. The best results of the drug users classifiers. Symbol 'X' means the used input feature. Results are calculated by LOOCV.

Target feature	Classifier	Age	Edu.	N	E	O	A	C	Imp.	SS	Gender	Sens. (%)	Spec. (%)	Sum (%)
Alcohol	LDA	X	X	X						X	X	75.34	63.24	138.58
Amphetamines	DT	X		X		X		X	X	X		81.30	71.48	152.77
Amyl nitrite	DT			X		X		X		X		73.51	87.86	161.37
Benzodiazepines	DT	X		X	X				X	X	X	70.87	71.51	142.38
Cannabis	DT	X	X			X	X	X	X			79.29	80.00	159.29
Chocolate	*k*NN	X			X			X			X	72.43	71.43	143.86
Cocaine	DT	X				X	X		X	X		68.27	83.06	151.32
Caffeine	*k*NN	X	X			X	X		X			70.51	72.97	143.48
Crack	DT				X			X				80.63	78.57	159.20
Ecstasy	DT	X								X	X	76.17	77.16	153.33
Heroin	DT	X							X		X	82.55	72.98	155.53
Ketamine	DT	X			X		X		X	X		72.29	80.98	153.26
Legal highs	DT	X				X	X	X		X	X	79.53	82.37	161.90
LSD	DT	X		X	X	X			X		X	85.46	77.56	163.02
Methadone	DT	X	X		X	X					X	79.14	72.48	151.62
MMushrooms	DT				X						X	65.56	94.79	160.36
Nicotine	DT			X	X			X			X	71.28	79.07	150.35
VSA	DT	X	X		X		X	X		X		83.48	77.64	161.12

Results

My results aren’t very different from those of the searchers. Generally, I obtained the best accuracy for almost the same combination of features except few features.

We do not have access to the hyperparameters they used in the studied so I used a grid search algorithm to tune the hyperparameters.

For all drugs except nicotine, the accuracy obtained on test sets is greater than 80% (70% for nicotine). It is without a doubt better than a random guess. We can infer that our personality traits play a huge role in our tendency to addiction.

In general random forest achieves the best accuracy.

Limits

For some drug we have few users, hence the accuracy for those one are biased by the dataset.

Results

	logistic regression	Classification tree	Random Forest	Boosting model	KNN	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS
accuracy for Alcohol	0.95491	0.95491	0.96021	0.95491	0.95491	X	X	X	X	X	X	X	X	X	X	X	X
Best parameters for model	{'C': 1}	{'criterion': 'entropy', 'max_depth': 3, 'max_...	{'criterion': 'entropy', 'max_depth': 5, 'max_...	{'max_depth': 5, 'max_features': 'log2', 'n_es...	{'n_neighbors': 20}												
accuracy for Amphet	0.80371	0.79576	0.81698	0.79576	0.81432	X		X	X	X	X	X	X	X	X	X	X
Best parameters for model	{'C': 1}	{'criterion': 'gini', 'max_depth': 3, 'max_fea...	{'criterion': 'gini', 'max_depth': 2, 'max_fea...	{'max_depth': 5, 'max_features': 'auto', 'n_es...	{'n_neighbors': 30}												
accuracy for Amyl	0.93634	0.93634	0.93634	0.93634	0.93634	X		X	X	X	X		X		X	X	X
Best parameters for model	{'C': 1}	{'criterion': 'entropy', 'max_depth': 2, 'max_...	{'criterion': 'entropy', 'max_depth': 2, 'max_...	{'max_depth': 5, 'max_features': 'log2', 'n_es...	{'n_neighbors': 20}												
accuracy for Benzos	0.72679	0.73475	0.75066	0.73475	0.7374	X	X		X	X	X	X		X		X	X
Best parameters for model	{'C': 1}	{'criterion': 'entropy', 'max_depth': 5, 'max_...	{'criterion': 'gini', 'max_depth': 5, 'max_fea...	{'max_depth': 5, 'max_features': 'auto', 'n_es...	{'n_neighbors': 70}												
accuracy for Caffeine	0.97082	0.97082	0.97082	0.97082	0.97082	X		X	X	X			X	X		X	

Here is a part of the best results obtained for all drugs and for all models after the gridsearch.