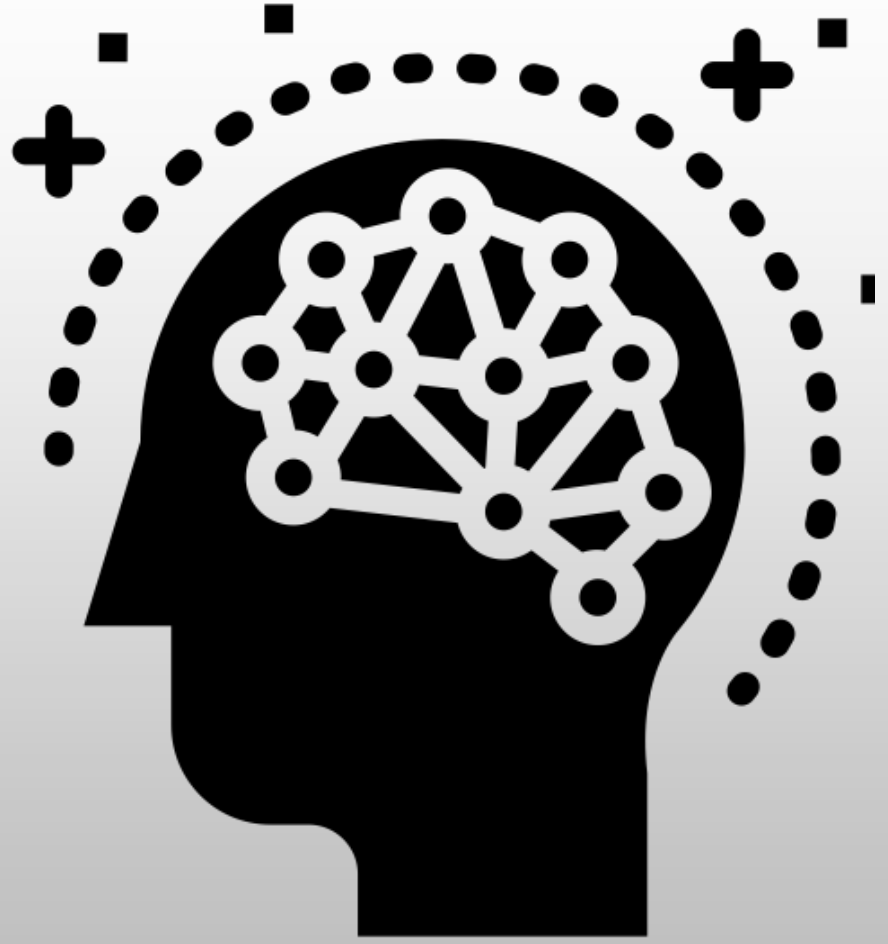


PYTHON FOR DATA ANALYSIS

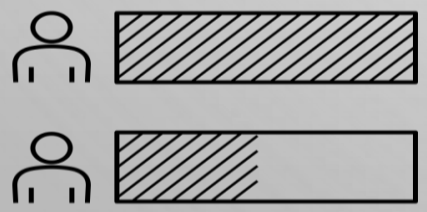
STUDY OF THE DRUG CONSUMPTION (QUANTIFIED) DATA SET

Anaïs Druart

THE DATASET

- 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
    - conscientiousness
  - ❑ BIS-11 (impulsivity)
  - ❑ impss (sensation seeking)
  - ❑ level of education
  - ❑ age
  - ❑ gender
  - ❑ country of residence
  - ❑ ethnicity



- All input attributes were originally categorical and were **quantified**
- 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
  - ❑ volatile substance abuse
  - ❑ Semeron

- For each drug they had to select one of the following answers:

  - ❑ Never used
  - ❑ Used over a decade ago
  - ❑ Used in the last decade
  - ❑ Used in the last year
  - ❑ Used in the last month
  - ❑ Used in the week
  - ❑ Used in the last day

PROBLEM

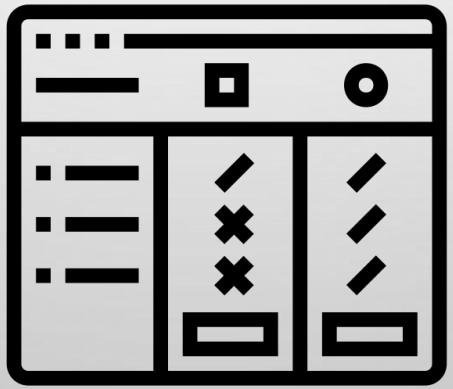
We have 12 attributes evaluated for each person, and a very large variety of drug (from common stimulants like chocolate and coffee, to strong drugs like LSD and heroin)

We want to produce **a model able to predict if a person could be a drug consumer**, for each type of drug

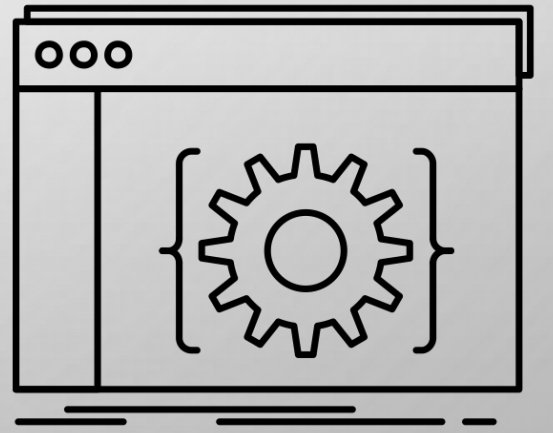
MAIN STEPS

EXPLORATORY DATA ANALYSIS (EDA) OF THE DATASET

MODEL CREATION, COMPARISON AND SELECTION



TRANSFORMATION OF THE MODEL TO AN API



Exploratory data analysis (EDA) of the dataset

Data-visualisation (with matplotlib, searborn, bokeh…)
Look for the link between the variables and the **target**

Checks on the dataset

```
data.head()
```

	ID	Age	Gender	Education	Country	Ethnicity	Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	Impulsiveness	Sensatio seeking
0	1	0.49788	0.48246	-0.05921	0.96082	0.12600	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1.18084
1	2	-0.07854	-0.48246	1.98437	0.96082	-0.31685	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575
2	3	0.49788	-0.48246	-0.05921	0.96082	-0.31685	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0.40148
3	4	-0.95197	0.48246	1.16365	0.96082	-0.31685	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084
4	5	0.49788	0.48246	1.98437	0.96082	-0.31685	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	-0.21575

Our dataset is far from being easy to read and understand. Let's make it more clear and convert nominal drug values into ordered data.

```
for i in drugs_columns:
    data[i] = data[i].map({'CL0': 0, 'CL1': 1, 'CL2': 2, 'CL3': 3, 'CL4': 4, 'CL5': 5, 'CL6': 6})
```

```
data.head()
```

nsation eking	Alcohol consumption	Amphetamines consumption	Amyl nitrite consumption	Benzodiazepine consumption	Caffeine consumption	Cannabis consumption	Chocolate consumption	Cocaine consumption	Crack consumption	Ecstasy consumption	H c
18084	5	2	0	2	6	0	5	0	0	0	0
21575	5	2	2	0	6	4	6	3	0	4	0
0148	6	0	0	0	6	3	4	0	0	0	0
18084	4	0	0	3	5	2	4	2	0	0	0
21575	4	1	1	0	6	3	6	0	0	1	0

The dataset doesn't have any missing values.

We added the **name of each column** according to the description of the dataset, and converted the nominal drug values into **ordered data**.

```
semerons = data[data['Fictitious drug Semeron consumption'] != 0]
semerons
```

	ID	Age	Gender	Education	Country	Ethnicity	Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	Impulsiveness	S
727	730	-0.07854	0.48246	-1.73790	-0.09765	-0.31685	-0.58016	0.32197	0.14143	-0.60633	0.12331	1.29221	
817	821	-0.95197	-0.48246	-0.61113	-0.09765	-0.50212	-0.67825	1.74091	0.72330	0.13136	0.41594	0.88113	
1516	1520	-0.95197	-0.48246	-0.61113	-0.57009	-0.31685	-0.24649	-0.80615	-1.27553	-1.34289	-1.92173	-0.71126	
1533	1537	-0.95197	0.48246	-0.61113	-0.57009	0.11440	-0.46725	0.80523	0.29338	2.03972	1.81175	-1.37983	
1698	1702	0.49788	0.48246	0.45468	-0.57009	-0.31685	1.98437	-0.80615	2.15324	0.76096	-0.00665	1.29221	
1769	1773	-0.95197	-0.48246	-1.22751	-0.57009	-0.22166	-0.34799	1.28610	1.06238	-0.01729	-0.52745	0.52975	
1806	1810	-0.95197	0.48246	-1.43719	-0.57009	-0.31685	1.23461	1.11406	1.06238	-1.47955	0.12331	0.88113	
1823	1827	-0.95197	0.48246	0.45468	-0.57009	-0.31685	0.22393	-0.30033	0.88309	1.28610	-0.00665	0.88113	

Semeron is a **fictitious drug** which was introduced to identify over-claimers. We exclude their records from further analysis.

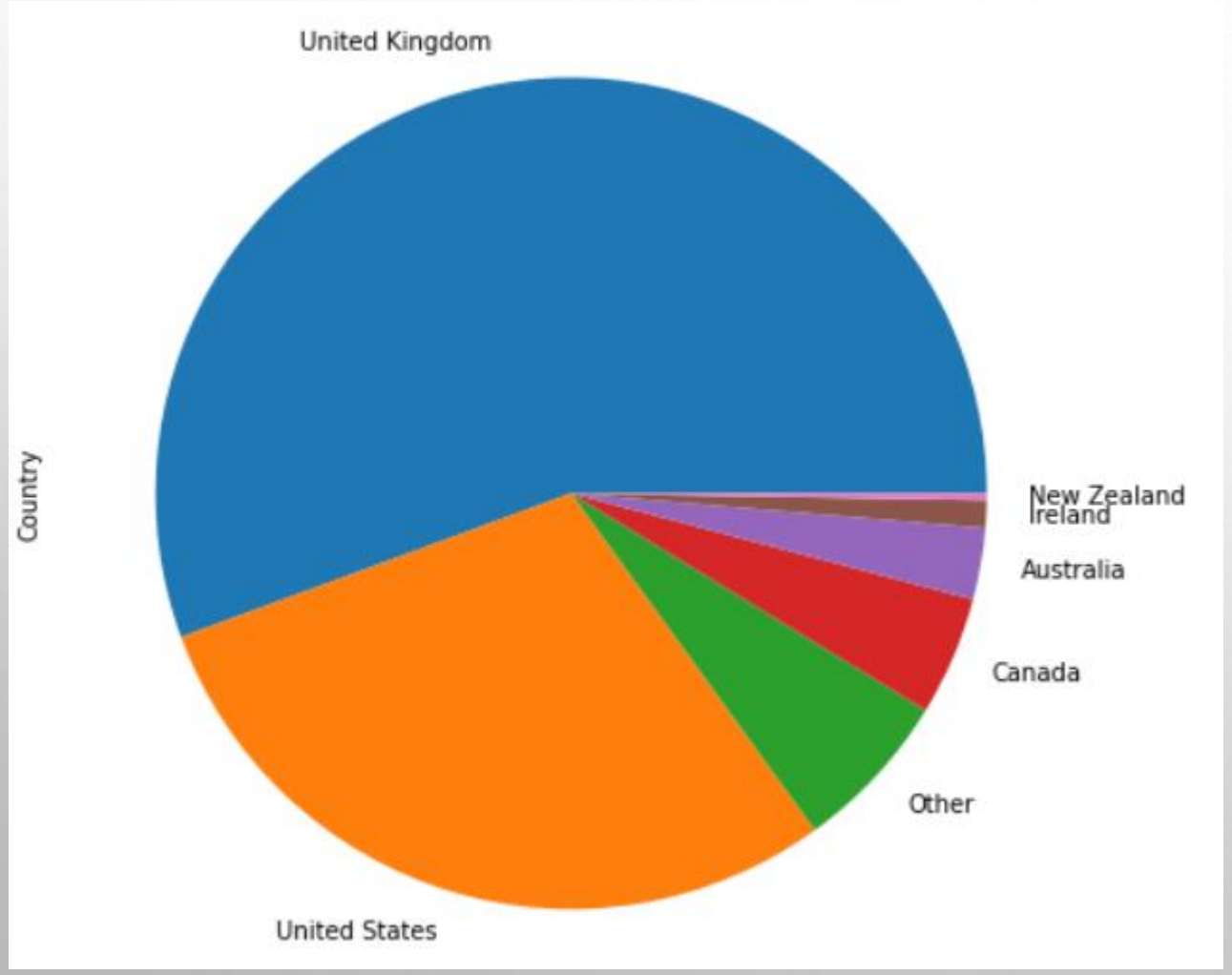
```
data_str.head()
```

	ID	Age	Gender	Education	Country	Ethnicity	Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	Impulsiveness	Sensatior seekinç
0	1	35-44	Female	Professional certificate/ diploma	United Kingdom	Mixed-White/Asian	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1.1808
1	2	25-34	Male	Doctorate degree	United Kingdom	White	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.2157
2	3	35-44	Male	Professional certificate/ diploma	United Kingdom	White	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0.4014
3	4	18-24	Female	Masters degree	United Kingdom	White	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.1808
4	5	35-44	Female	Doctorate degree	United Kingdom	White	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	-0.2157

Way more understandable !

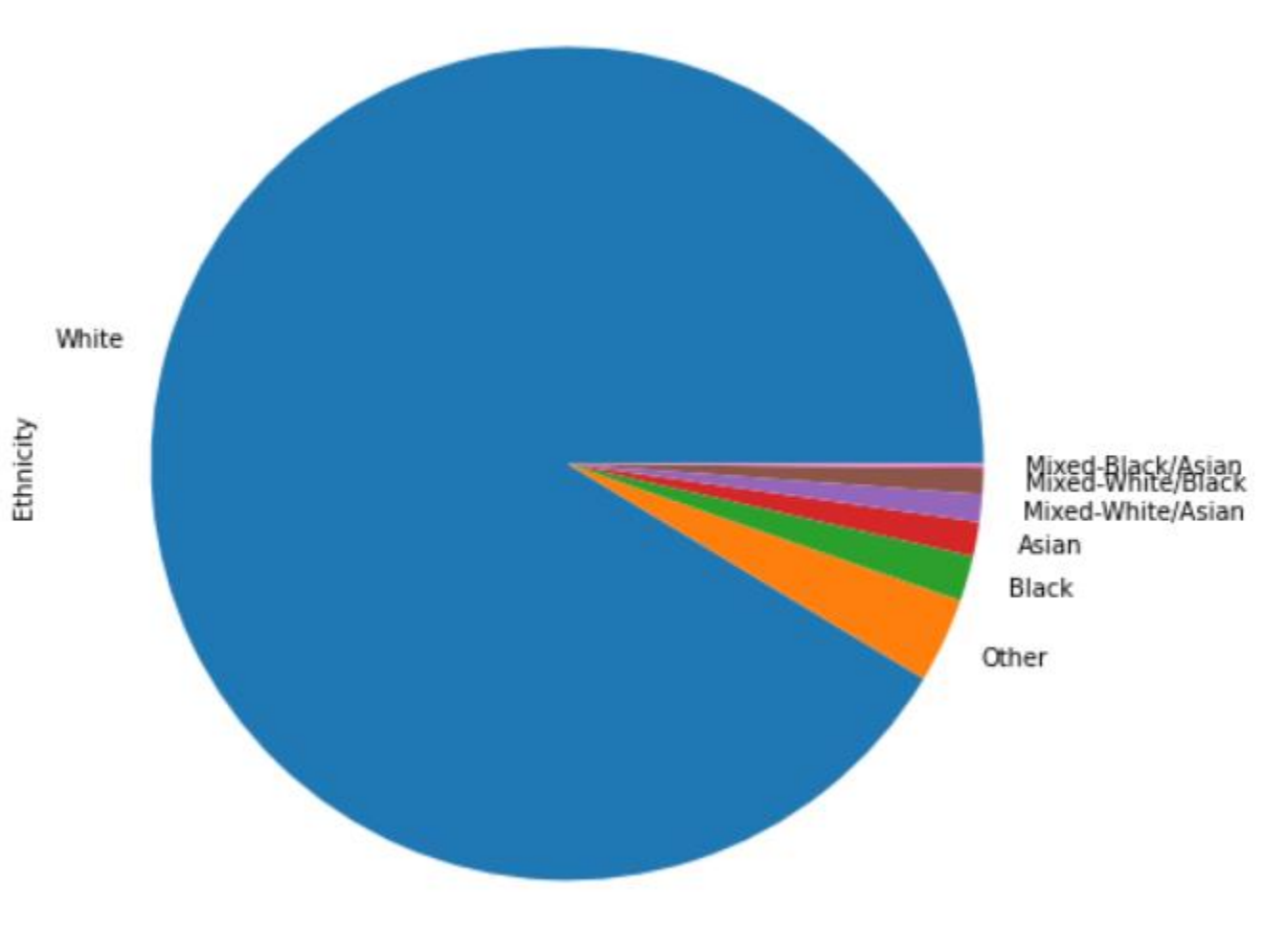
As we can see, the dataset still needed more cleaning. All input attributes were originally categorical and then quantified, so for analysis purposes it was retrieved to a more understandable and clearer form.

Distribution of the respondents per country



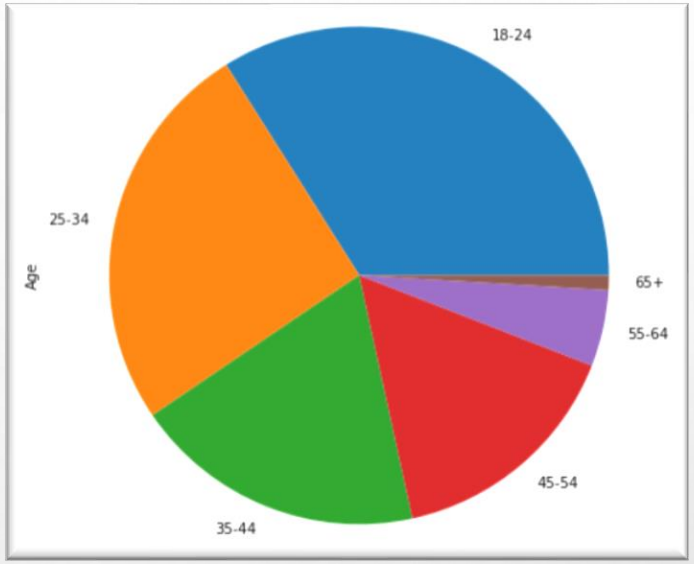
Most of the participants comes from the **UK and the USA**. All countries are officially English-speaking. It's **non representative** of the entirety of the world.

Distribution of the respondents per ethnicity



Non-surprisingly, more than 90% of the participants are white. It's an **important bias** to consider.

Distribution of the respondents per age

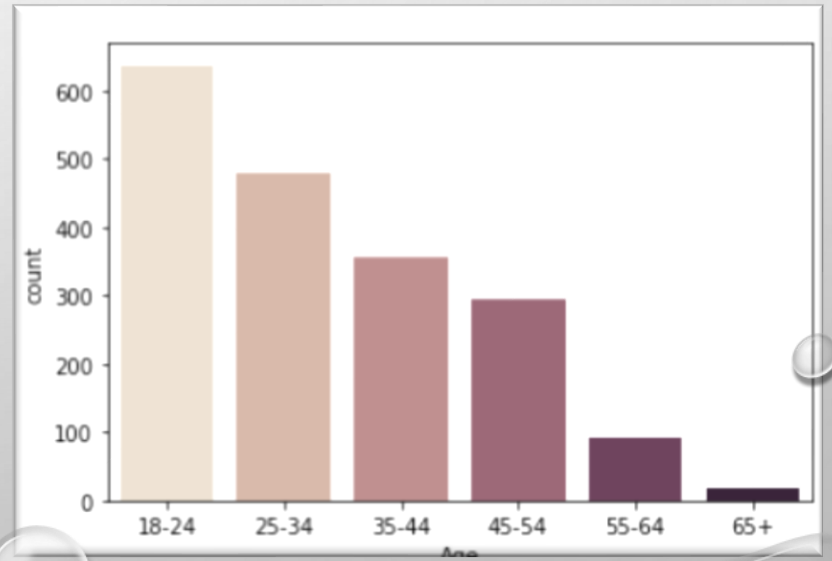


Age (Real) is the age of the participants.

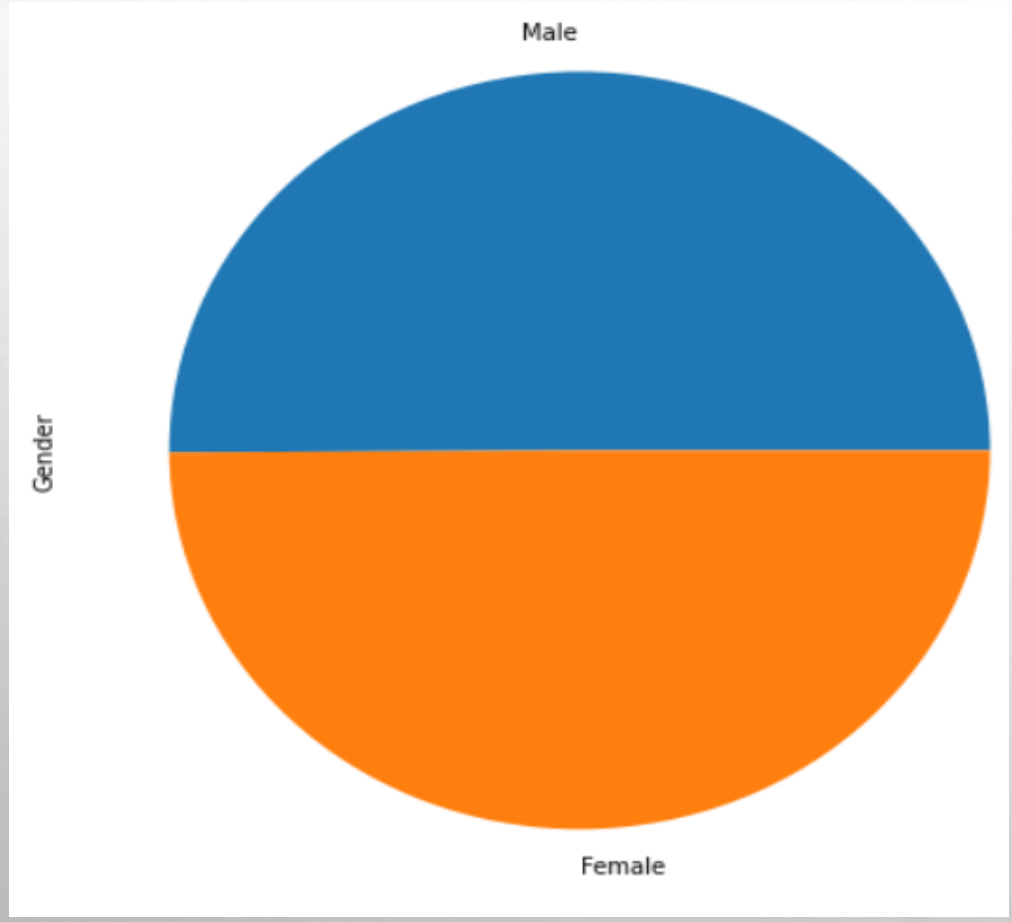
The age is given in the intervals of about 8-10 years.

18-24 is the biggest age group (about 1/3 of all participants).

25-34 is 1/4 of participants. 35-54 is another 1/3 of participants. The rest 5% are the people above 55 y.o.



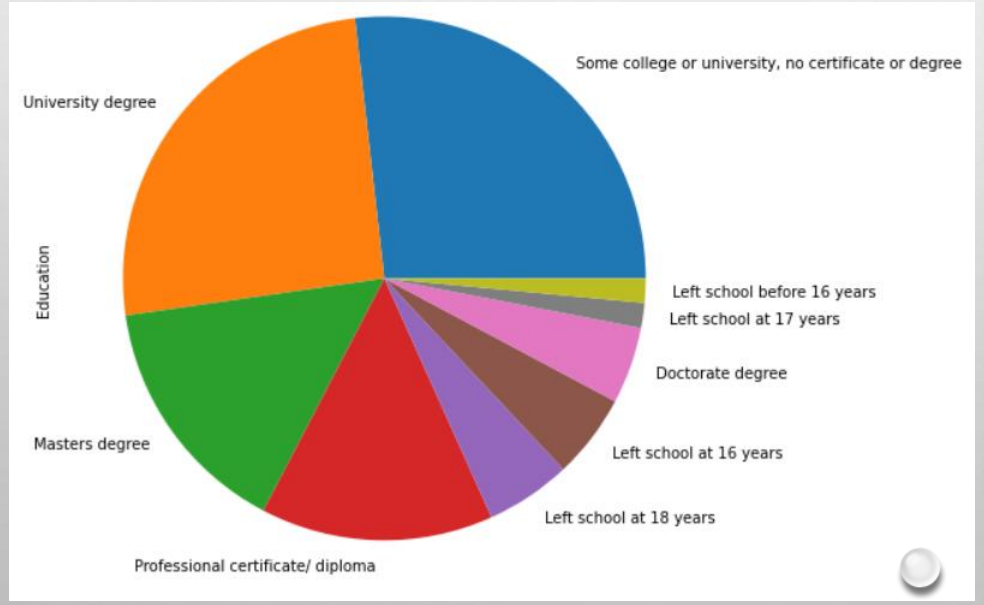
Distribution of the respondents per gender

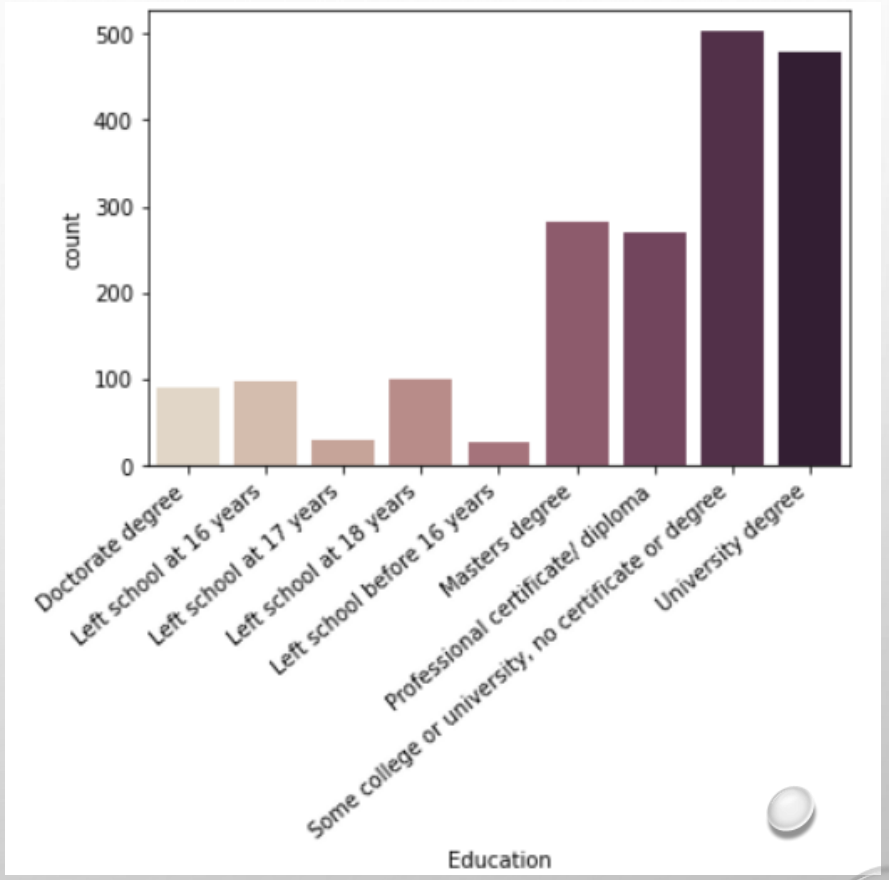


There is nearly the **same proportion** of male and female participants.

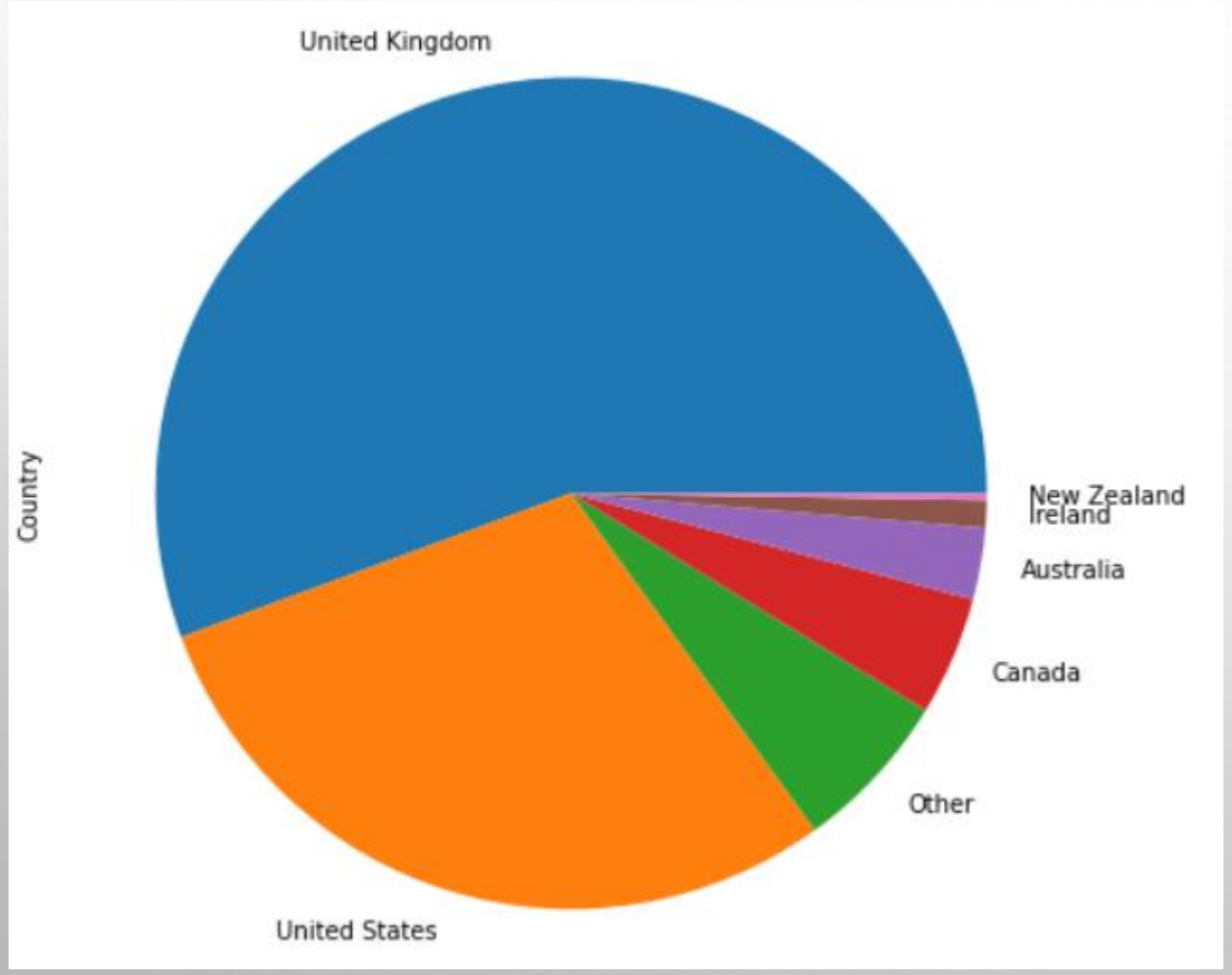
Distribution of the respondents per education

The big majority of the participants **is well educated** and has a university degree or some college experience.



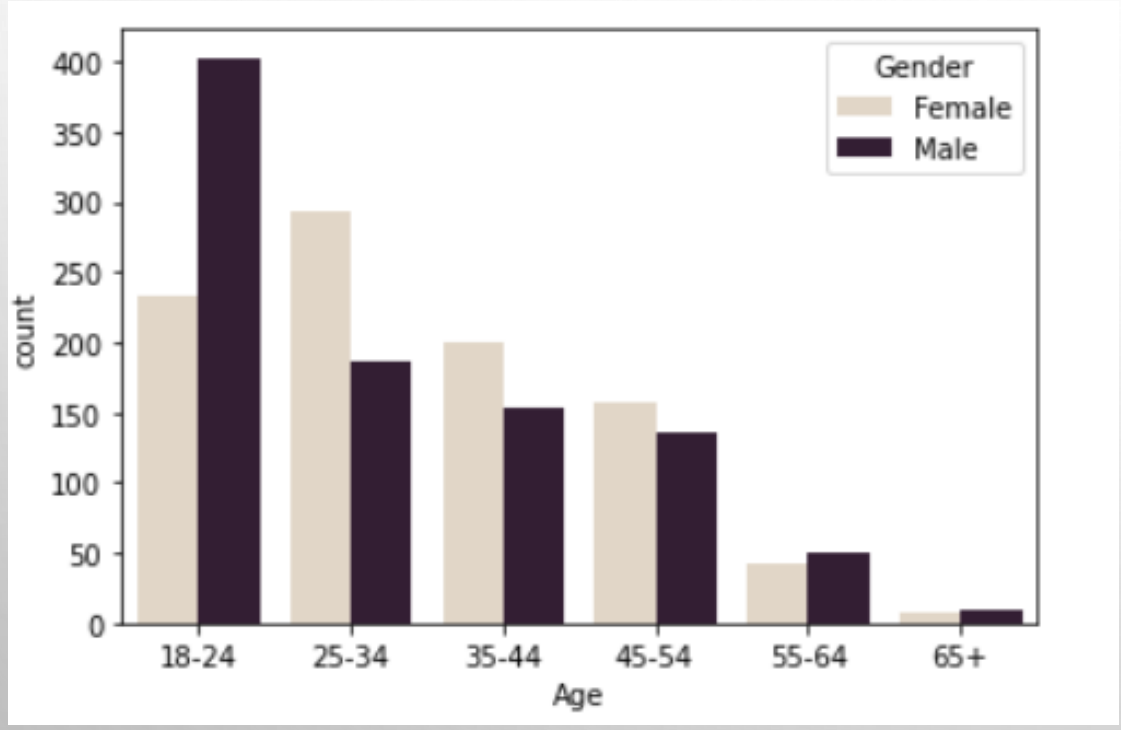


Distribution of the respondents per country



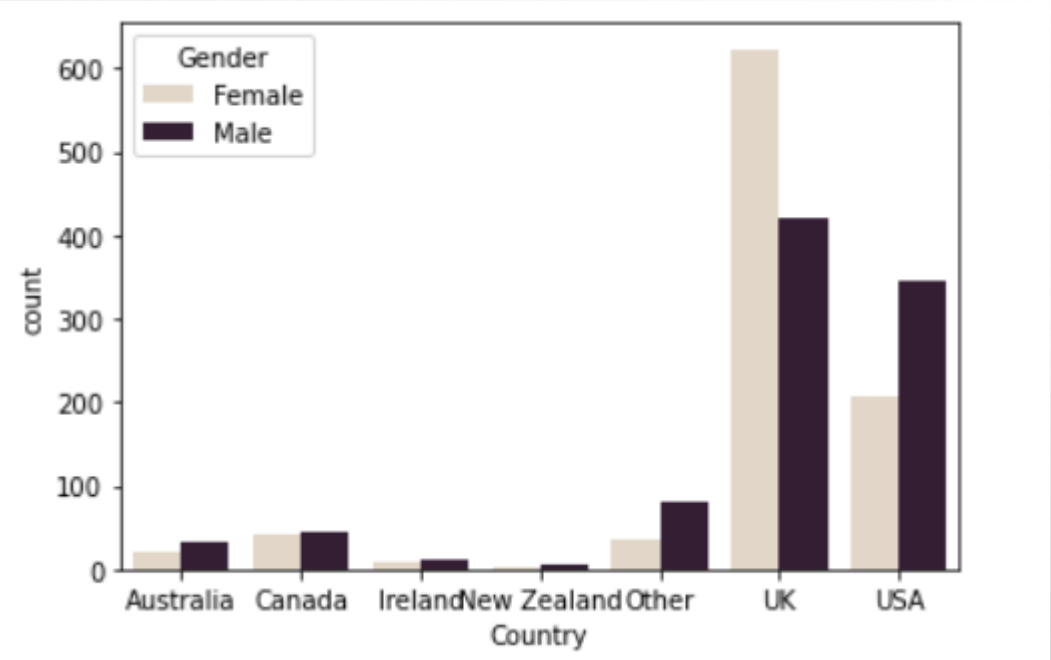
Most of the participants comes from the **UK and the USA**. All countries are officially English-speaking. It's **non representative** of the entirety of the world.

Age-gender distribution



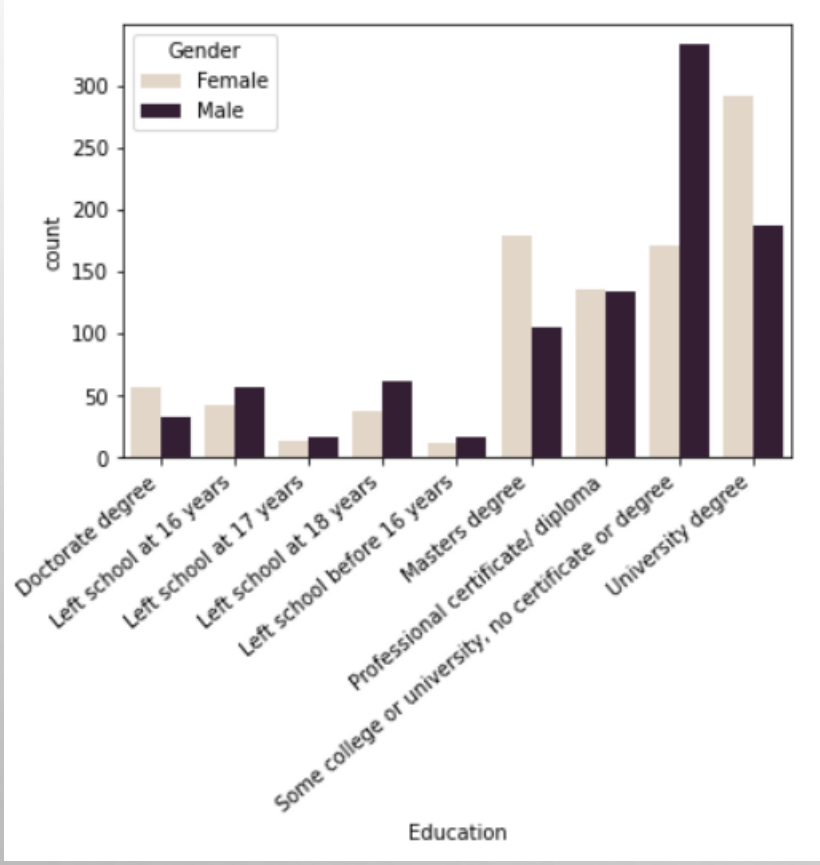
The males are predominant in the age group 18-24.
Then the females predominate decreasingly inf the age groups 25-34, 35-44, 45-54.
Finally, the males predominate very slightly for the groups higher than 55 years old.

Country-gender distribution



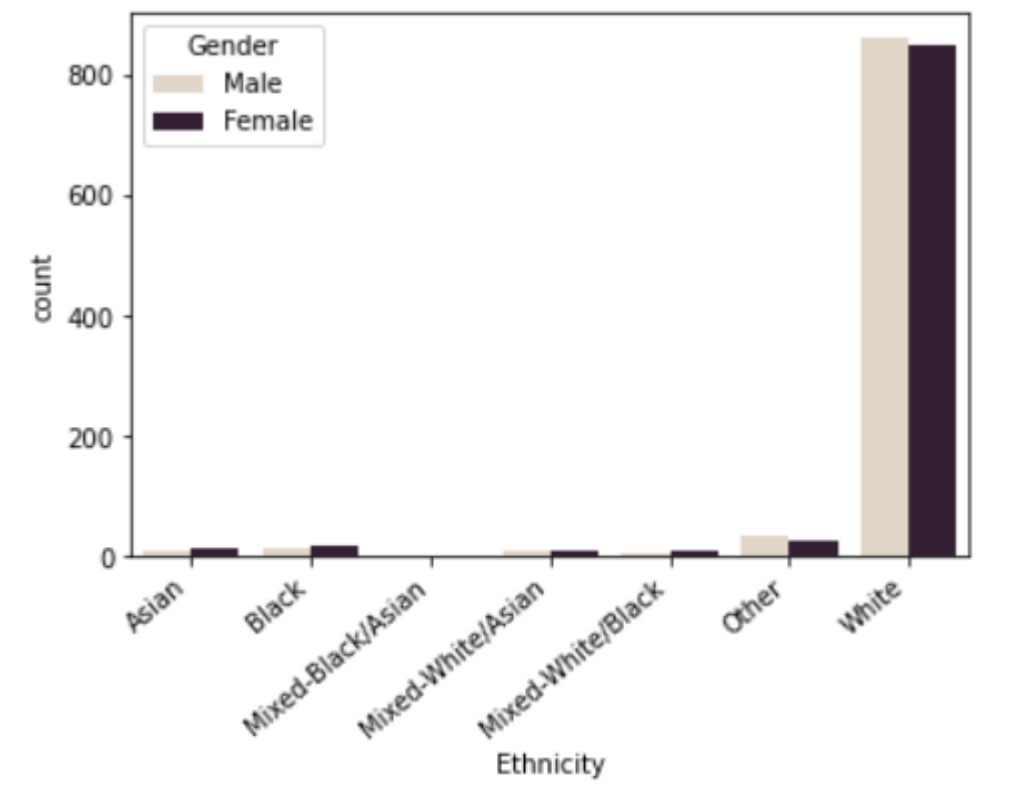
More women were tested in the UK and more men were tested in the USA.

Education-gender distribution



The females group predominate in the University degrees and master degrees, while the males group predominate in only some college experience. For the rest, the two groupes are kind of equally represented.

Ethnicity-gender distribution



The males and females groups are both equally represented in each ethnicity.

Age-gender-country crosstable

Gender	Female							Male						
Country	Australia	Canada	Ireland	New Zealand	Other	United Kingdom	United States	Australia	Canada	Ireland	New Zealand	Other	United Kingdom	United States
Age														
18-24	1	7	3	0	12	112	99	20	25	5	2	39	91	221
25-34	10	9	2	1	13	200	58	5	8	3	2	26	80	63
35-44	4	8	3	0	7	150	29	3	7	0	0	10	108	26
45-54	5	7	1	0	4	127	14	3	4	3	0	5	99	22
55-64	0	7	0	0	0	29	7	1	1	0	0	2	36	10
65+	0	3	0	0	0	5	0	0	1	0	0	0	7	2

Only one male over 54 years old was tested in Australia, while no people over 54 years old were tested in Ireland, and no people over 34 years old were tested in New Zealand.
Way more males of 18-24 years old than females were tested in the USA.

Education-gender-ethnicity crosstable

Gender	Female							Male						
Ethnicity	Asian	Black	Mixed-Black/Asian	Mixed-White/Asian	Mixed-White/Black	Other	White	Asian	Black	Mixed-Black/Asian	Mixed-White/Asian	Mixed-White/Black	Other	White
Education														
Doctorate degree	1	0	0	1	2	1	52	1	0	0	0	0	2	29
Left school at 16 years	0	0	0	0	0	2	40	0	0	0	1	1	0	54
Left school at 17 years	0	0	0	0	1	0	12	0	0	0	0	0	2	14
Left school at 18 years	0	0	0	0	0	2	35	0	1	0	1	2	2	56
Left school before 16 years	0	0	0	0	0	0	12	0	0	0	0	0	0	16
Masters degree	4	6	0	3	5	5	156	5	1	0	0	1	4	93
Professional certificate/ diploma	0	3	0	1	0	2	130	1	5	0	1	1	6	120
Some college or university, no certificate or degree	3	2	1	2	1	8	153	0	4	1	4	3	12	309
University degree	7	6	1	4	2	8	263	3	5	0	2	0	6	171

Country-gender-ethnicity crosstable

Gender	Female							Male						
Ethnicity	Asian	Black	Mixed-Black/Asian	Mixed-White/Asian	Mixed-White/Black	Other	White	Asian	Black	Mixed-Black/Asian	Mixed-White/Asian	Mixed-White/Black	Other	White
Country														
Australia	0	0	0	0	0	0	20	0	0	0	0	0	0	32
Canada	0	0	0	2	0	1	38	0	0	0	0	2	2	42
Ireland	0	0	0	0	0	0	9	0	0	0	0	0	0	11
New Zealand	0	0	0	0	0	0	1	0	0	0	0	0	0	4
Other	1	1	0	0	2	2	30	1	0	1	0	1	5	74
UK	11	14	0	6	9	12	571	9	11	0	2	1	4	394
USA	3	2	2	3	0	13	184	0	5	0	7	4	23	305

Overview of the consumption of the drugs

	0	1	2	3	4	5	6
Alcohol	33	34	68	197	284	758	503
Amphetamines	973	230	241	196	75	61	101
Amyl nitrite	1299	210	236	91	24	14	3
Benzodiazepine	999	116	230	234	119	84	95
Caffeine	27	10	24	59	106	271	1380
Cannabis	413	207	266	210	138	185	458
Chocolate	32	2	10	53	295	680	805
Cocaine	1036	160	267	257	98	40	19
Crack	1622	67	109	59	9	9	2
Ecstasy	1020	112	232	275	154	63	21
Heroin	1600	68	91	65	24	16	13
Ketamine	1488	43	140	129	40	33	4
Legal highs	1092	29	195	321	109	64	67
LSD	1069	257	175	213	96	55	12
Methadone	1424	39	95	148	50	48	73
Magic mushrooms	982	208	259	272	114	39	3
Nicotine	428	193	203	184	106	156	607
VSA	1452	199	133	59	13	14	7

Alcohol, Caffeine, Chocolate and Nicotine are stimulants or treats that are consumed rather often so that's why we have **a vast majority of users.**

Cannabis **distributes evenly** : there is nearly the same number of people that took it lately as the number of people that never tried it.

Amphetamine, Amyl nitrite, Benzodiazepine, Cocaine, Ecstasy, Heroin, Ketamine, Legal highs, LSD, Methadone, Magic mushrooms, Volatile substance abuse - **the vast majority of people never tested it**, or did it long time ago rather than recently.

Personality measurements

	Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	Impulsiveness	Sensation seeking
count	1877.000000	1877.000000	1877.000000	1877.000000	1877.000000	1877.000000	1877.000000
mean	-0.000551	-0.001951	-0.003224	-0.000657	-0.000394	0.005293	-0.007408
std	0.998442	0.997418	0.995691	0.996689	0.997657	0.954148	0.962074
min	-3.464360	-3.273930	-3.273930	-3.464360	-3.464360	-2.555240	-2.078480
25%	-0.678250	-0.695090	-0.717270	-0.606330	-0.652530	-0.711260	-0.525930
50%	0.042570	0.003320	-0.019280	-0.017290	-0.006650	-0.217120	0.079870
75%	0.629670	0.637790	0.723300	0.760960	0.584890	0.529750	0.765400
max	3.273930	3.273930	2.901610	3.464360	3.464360	2.901610	1.921730

```
corr = pers_data.corr(method="spearman")
corr
```

	Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	Impulsiveness	Sensation seeking
Neuroticism	1.000000	-0.416843	0.015156	-0.205869	-0.377803	0.168010	0.090256
Extraversion	-0.416843	1.000000	0.226260	0.161784	0.290479	0.119210	0.189641
Openness to experience	0.015156	0.226260	1.000000	0.036995	-0.073167	0.270398	0.404787
Agreeableness	-0.205869	0.161784	0.036995	1.000000	0.236550	-0.225828	-0.208106
Conscientiousness	-0.377803	0.290479	-0.073167	0.236550	1.000000	-0.344500	-0.254003
Impulsiveness	0.168010	0.119210	0.270398	-0.225828	-0.344500	1.000000	0.628217
Sensation seeking	0.090256	0.189641	0.404787	-0.208106	-0.254003	0.628217	1.000000

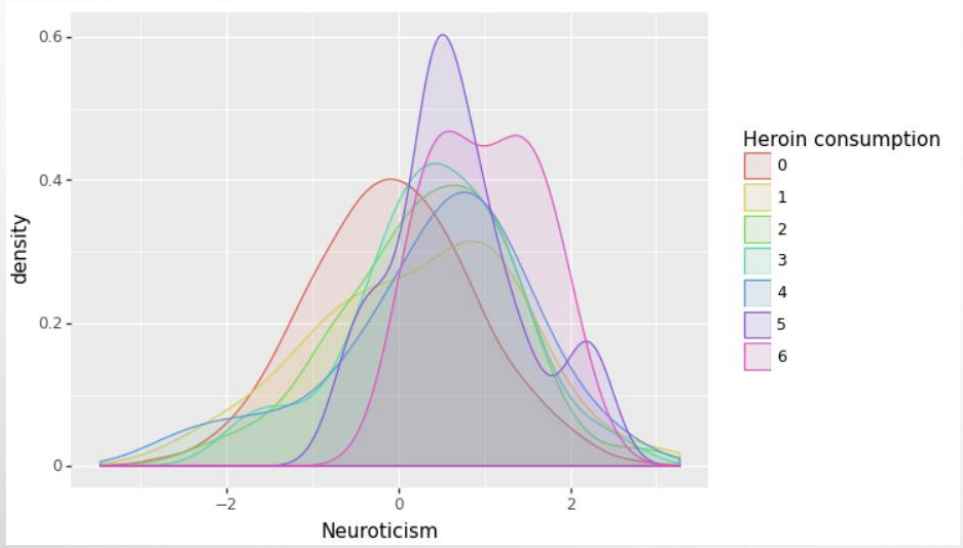
Personality measurements

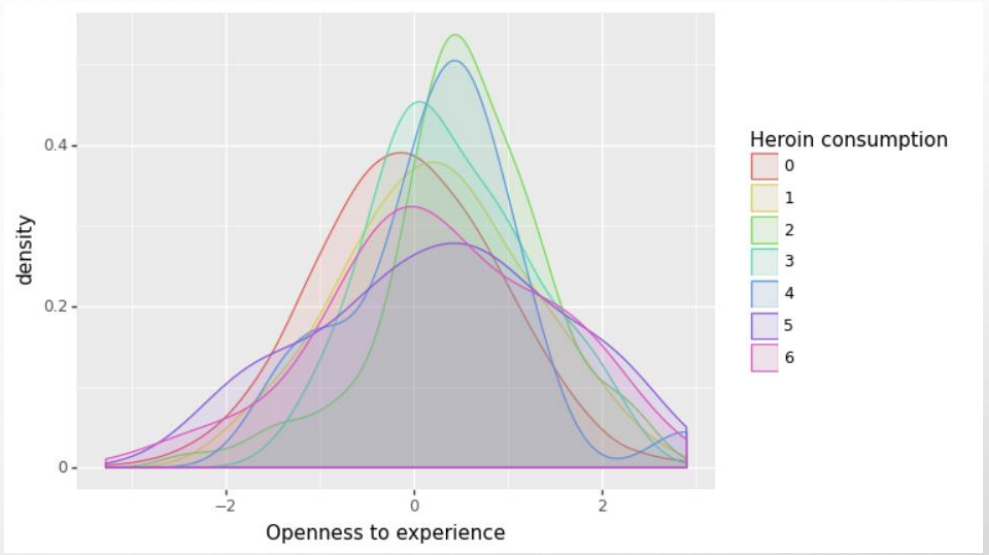
The correlations between personality factors are not very strong, **at most moderate**.

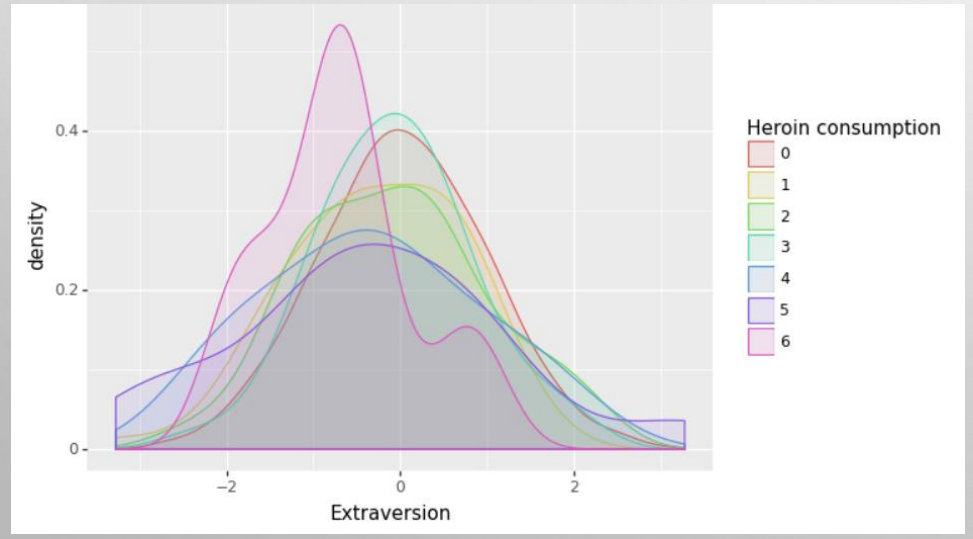
The strongest correlations positive:
Sensation seeking and Impulsiveness (r = 0.63)
Sensation seeking and Openness to experience (r = 0.4)

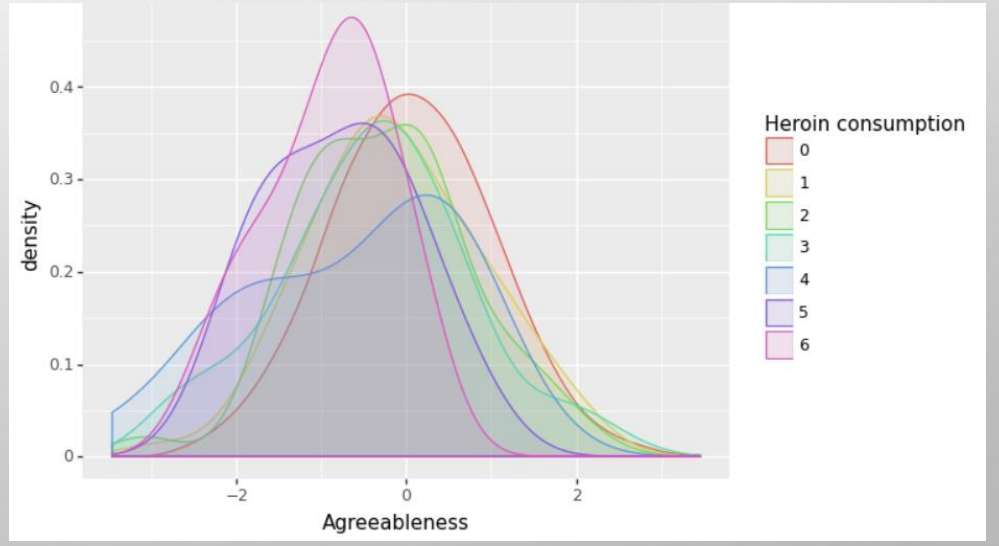
The strongest correlations negative:
Extraversion and Neuroticism (r = -0.42)
Neuroticism and Conscientiousness (r = -0.38)
Conscientiousness and Impulsiveness (r = -0.34)

Correlation between the heroin consumption and the personality measurements of a participant

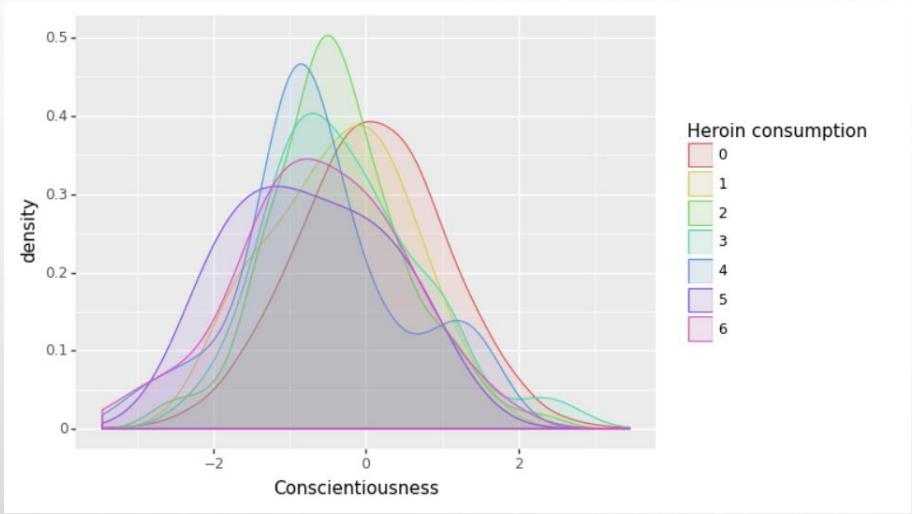


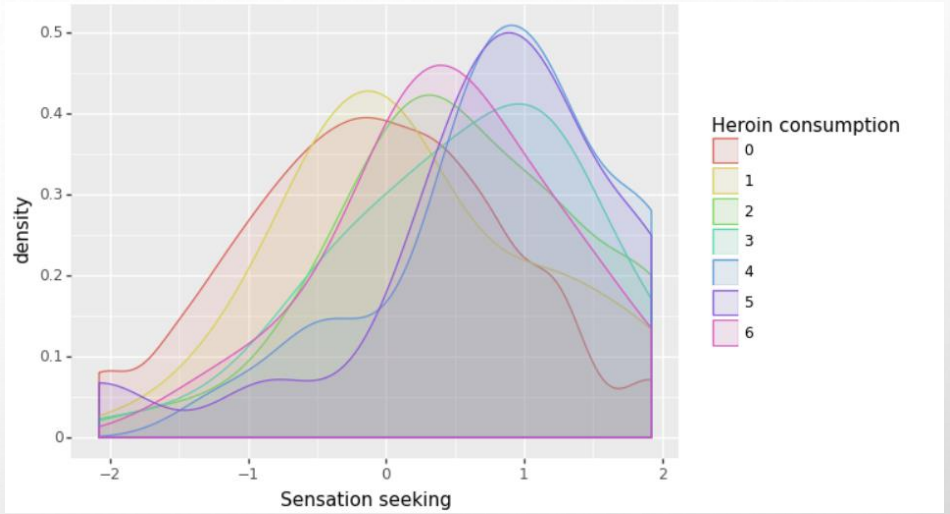


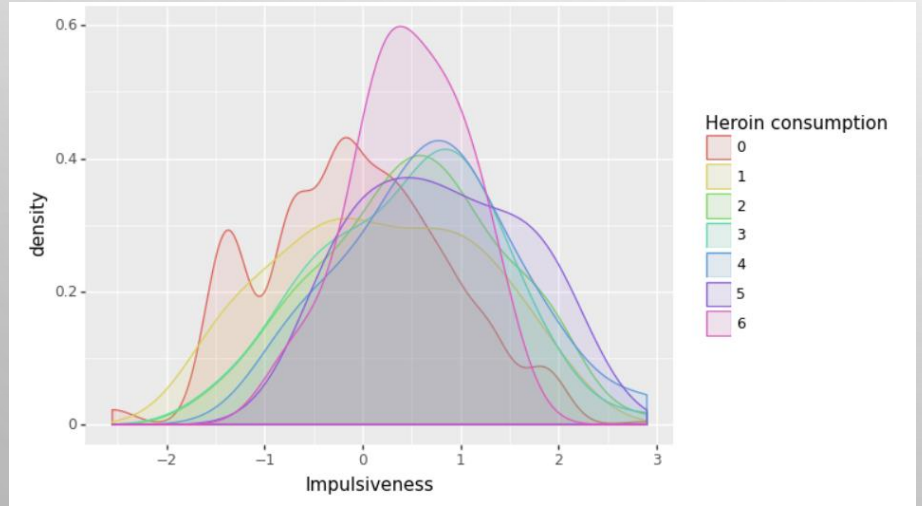




Correlation between the heroin consumption and the personality measurements of a participant – part 2







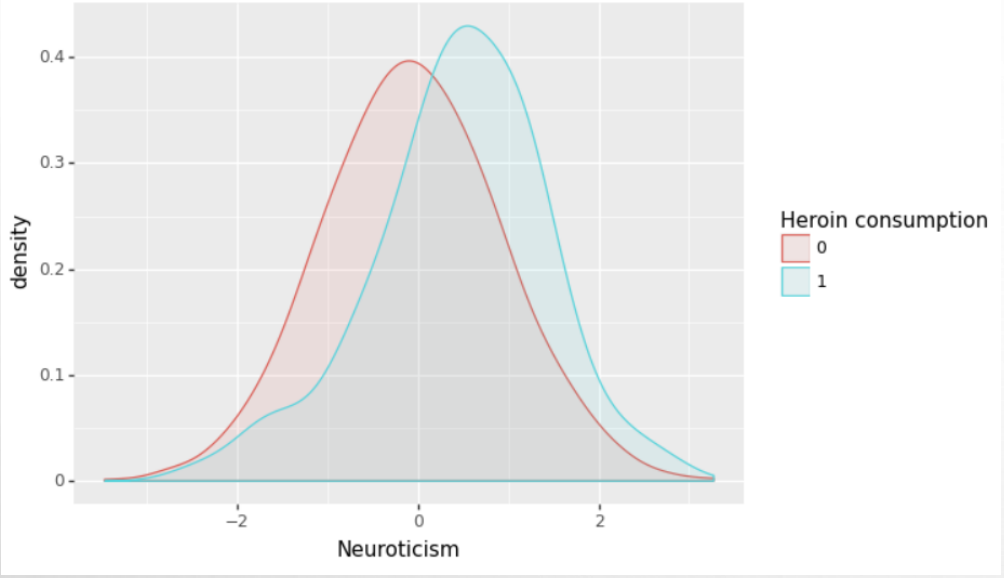
As we can see, the people consuming heroin tend to have more neuroses, to be more opened to new experiences, more impulsive and tend to search for more sensations; while they are less extravert, agreeable, or conscientious.

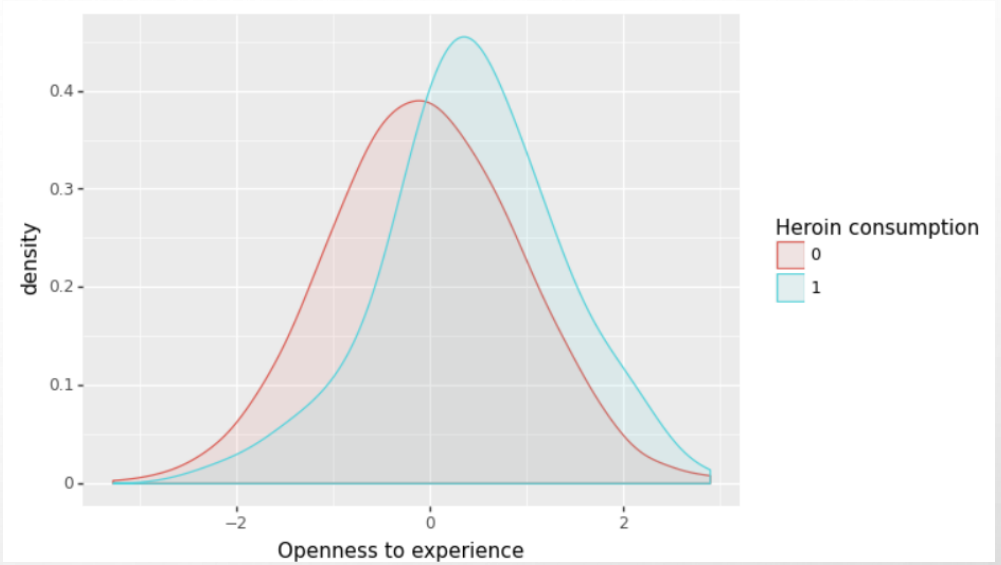
Binary classification by union of part of classes into one new class

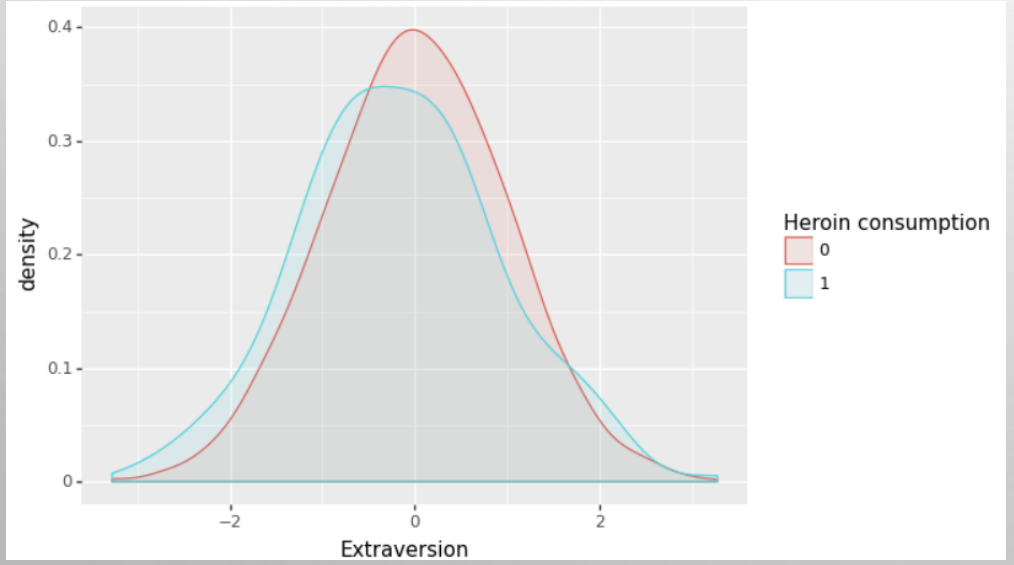
Alcohol nsumption	Amphetamines consumption	Amyl nitrite consumption	Benzodiazepine consumption	Caffeine consumption	Cannabis consumption	Chocolate consumption	Cocaine consumption	Crack consumption	Ecstasy consumption	Heroin consumption
1	1	0	1	1	0	1	0	0	0	0
1	1	1	0	1	1	1	1	0	1	0
1	0	0	0	1	1	1	0	0	0	0
1	0	0	1	1	1	1	1	0	0	0
1	0	0	0	1	1	1	0	0	0	0

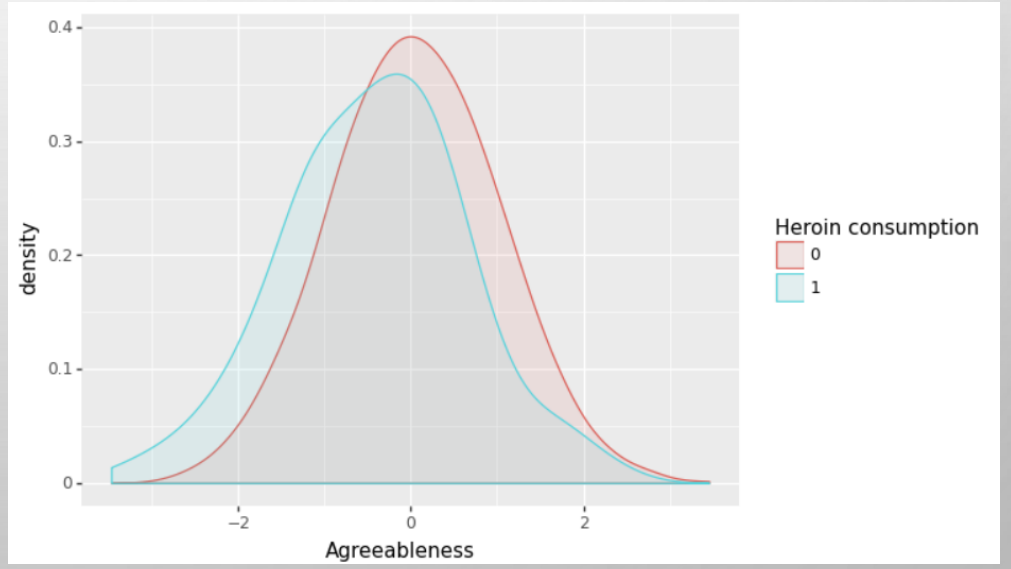
"Never used", "used over a decade ago" form class "non-user" and all other classes form class "user"

Correlation between the heroin consumption and the personality measurements of a participant, with only two classes

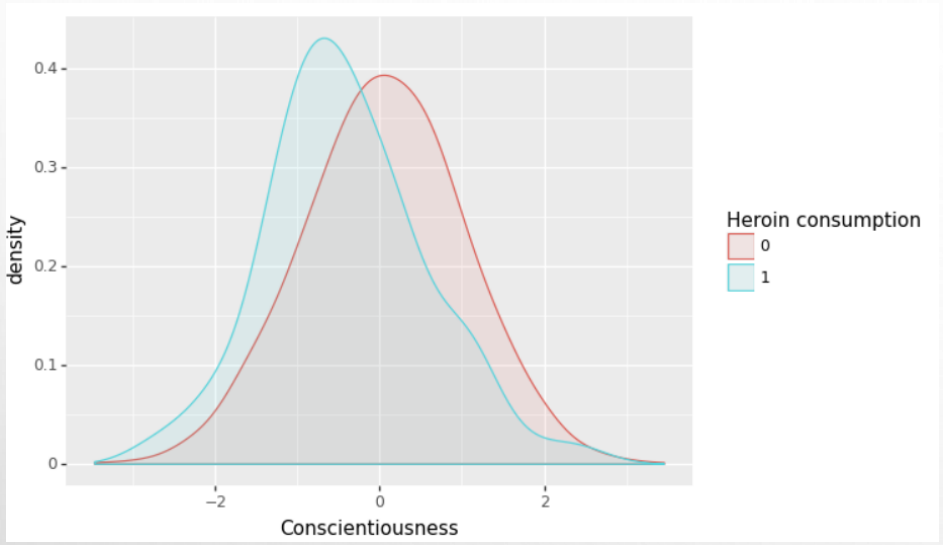


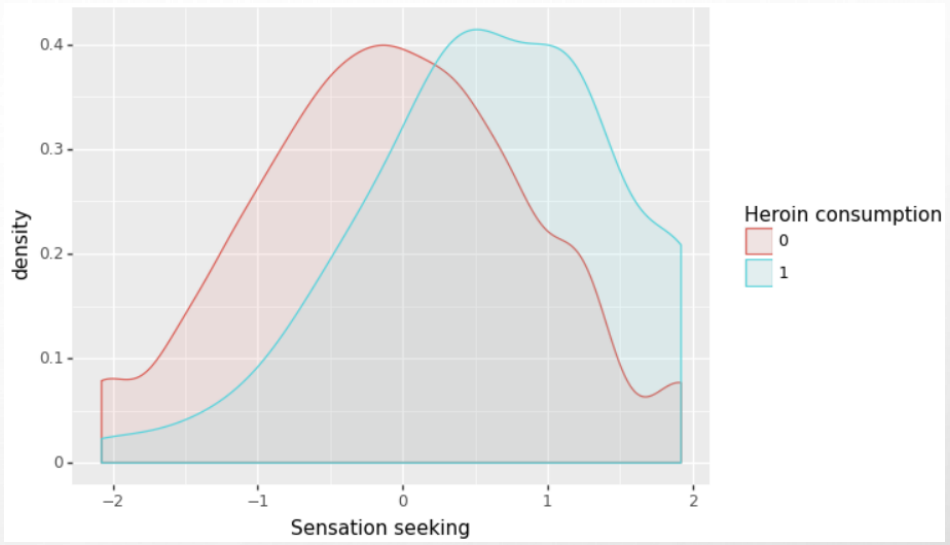


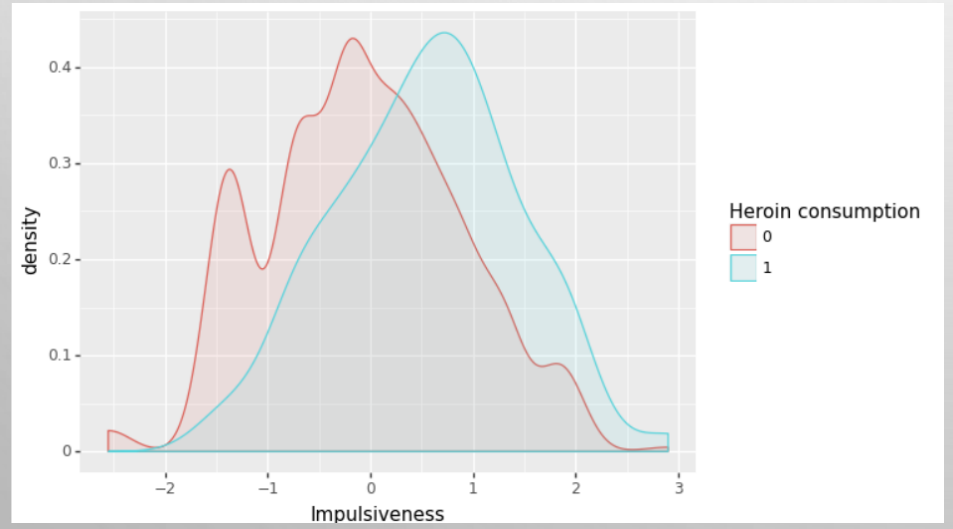




Correlation between the heroin consumption and the personality measurements of a participant, with only two classes – part 2





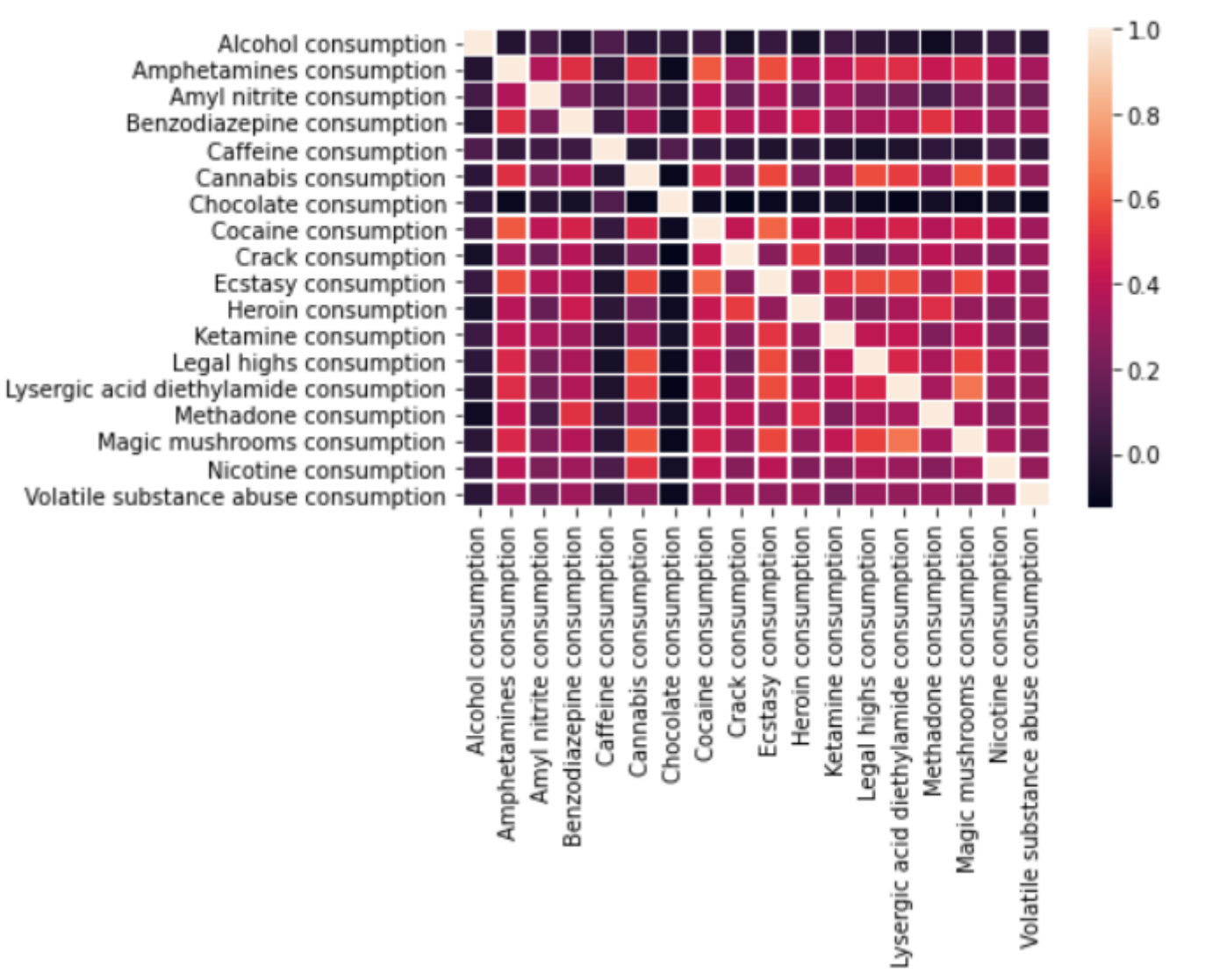


As we can see, **the results we obtain are globally the same.**

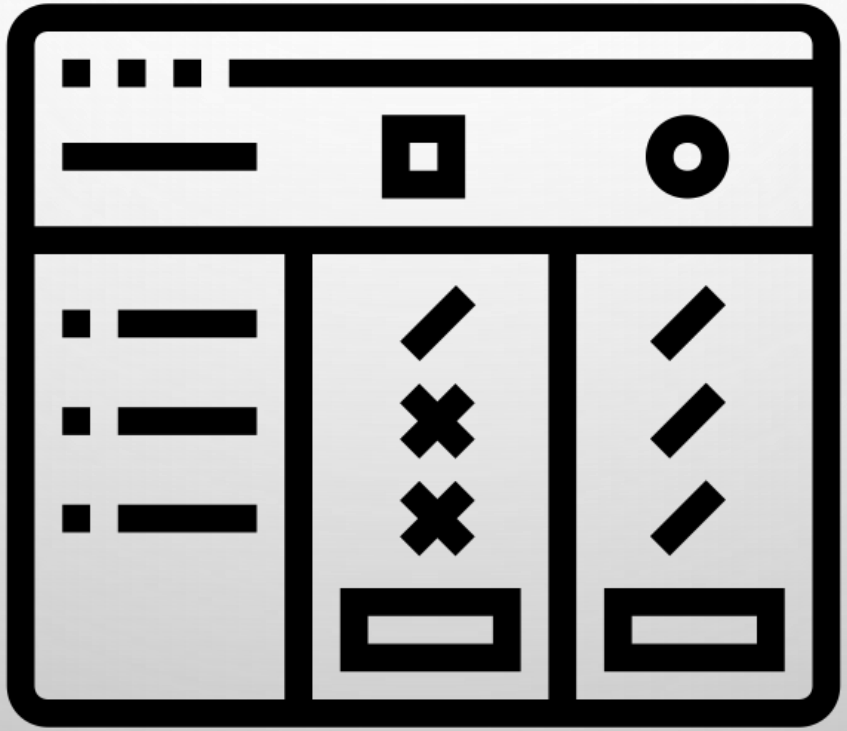
P-values : correlation between drugs consumption and personality traits

	Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	Impulsiveness	Sensation seeking
Alcohol consumption	0.0	0.084 **	0.032	-0.033	0.006	0.039	0.094 **
Amphetamines consumption	0.135 **	-0.042	0.253 **	-0.14 **	-0.253 **	0.294 **	0.365 **
Amyl nitrite consumption	0.04	0.04	0.065 **	-0.079 **	-0.121 **	0.13 **	0.189 **
Benzodiazepine consumption	0.266 **	-0.096 **	0.225 **	-0.164 **	-0.217 **	0.233 **	0.255 **
Caffeine consumption	0.022	0.013	-0.02	-0.008	0.006	0.014	-0.005
Cannabis consumption	0.11 **	-0.022	0.417 **	-0.155 **	-0.293 **	0.313 **	0.465 **
Chocolate consumption	0.025	0.028	-0.015	0.038	0.023	-0.015	-0.052 **
Cocaine consumption	0.144 **	0.006	0.205 **	-0.183 **	-0.221 **	0.264 **	0.34 **
Crack consumption	0.118 **	-0.052 **	0.126 **	-0.091 **	-0.13 **	0.19 **	0.197 **
Ecstasy consumption	0.087 **	0.056 **	0.311 **	-0.118 **	-0.249 **	0.272 **	0.405 **
Heroin consumption	0.178 **	-0.076 **	0.163 **	-0.14 **	-0.17 **	0.198 **	0.216 **
Ketamine consumption	0.078 **	0.019	0.193 **	-0.125 **	-0.167 **	0.188 **	0.266 **
Legal highs consumption	0.122 **	-0.033	0.347 **	-0.136 **	-0.27 **	0.281 **	0.441 **
Lysergic acid diethylamide consumption	0.071 **	-0.021	0.368 **	-0.12 **	-0.198 **	0.253 **	0.381 **
Methadone consumption	0.194 **	-0.108 **	0.209 **	-0.153 **	-0.209 **	0.198 **	0.242 **
Magic mushrooms consumption	0.066 **	0.001	0.379 **	-0.127 **	-0.221 **	0.281 **	0.391 **
Nicotine consumption	0.133 **	-0.033	0.193 **	-0.113 **	-0.237 **	0.253 **	0.306 **
Volatile substance abuse consumption	0.132 **	-0.066 **	0.151 **	-0.125 **	-0.182 **	0.192 **	0.255 **

Association rules



MODEL CREATION, COMPARISON AND SELECTION

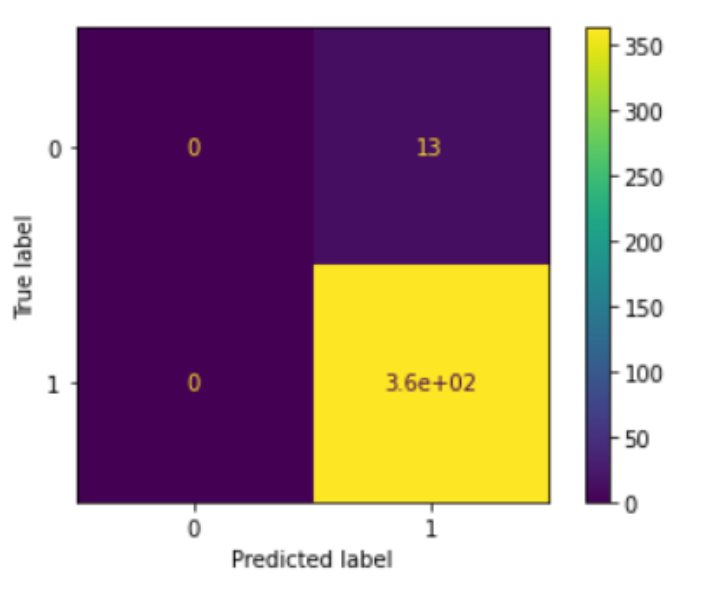


Alcohol consumption -models

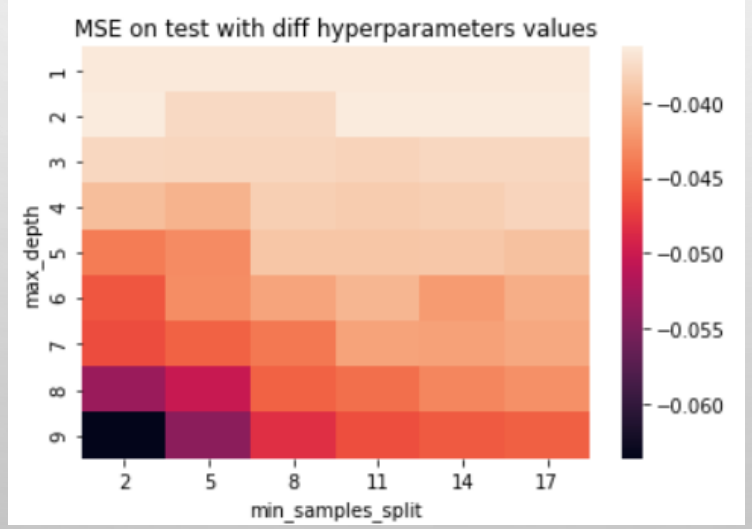
Encoding the categorical features

otine ption	Fictitious drug Semeron consumption	Volatile substance abuse consumption	has_25-34	has_35-44	has_45-54	has_55-64	has_65+	is_Male	studied_Left school at 16 years	studied_Left school at 17 years	studied_Left school at 18 years	studied_Left school before 16 years	studied_Mas deg
1	0	0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0	0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	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	
1	0	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0	0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

Alcohol consumption -models



Logistic regression model

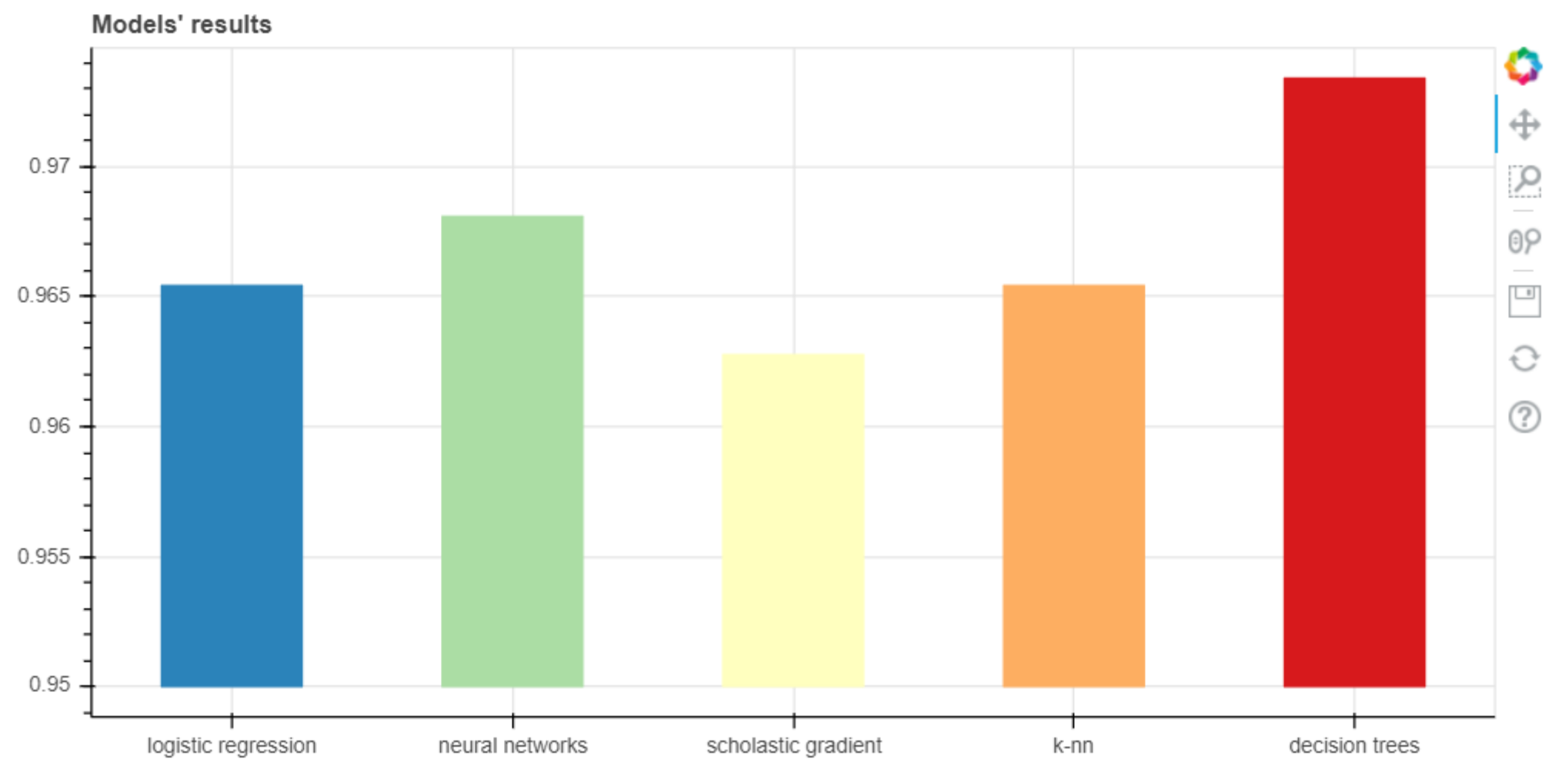


Decision trees model

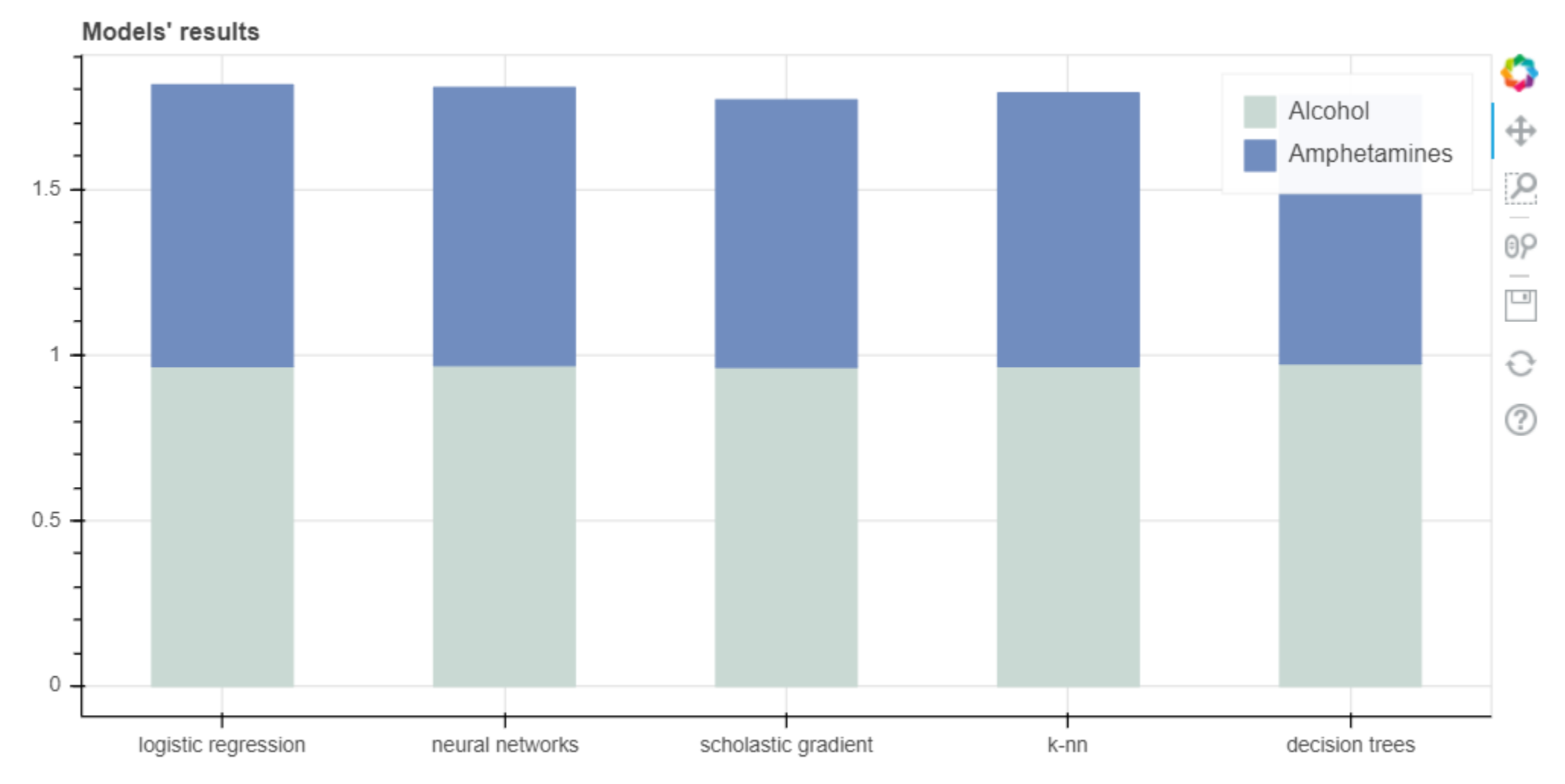
Alcohol consumption -models

Name of the model	Max Accuracy	Tuning
Logistic regression	96.5%	Yes : max_iter = 1000
Neural network	98.1%	Yes : i made a pipeline : make_pipeline(StandardScaler(), MLPClassifier(solver='adam', alpha=1e-5, hidden_layer_sizes=(5,2), random_state=1))
Stochastic gradient	97.6%	Yes : i made a pipeline
K Nearest Neighbors (K-NN)	96.8%	Yes : i used the Neighborhood Components Analysis with random_state = 42 and k=3
Decision trees	97.6%	Yes : i did a Gridsearch CV with KFold. tree.DecisionTreeClassifier(ccp_alpha=0.0, max_depth=2, max_features=**None**, max_leaf_nodes=**None**, min_impurity_decrease=0.0, min_impurity_split=**None**, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort='deprecated', random_state=**None**, splitter='best')

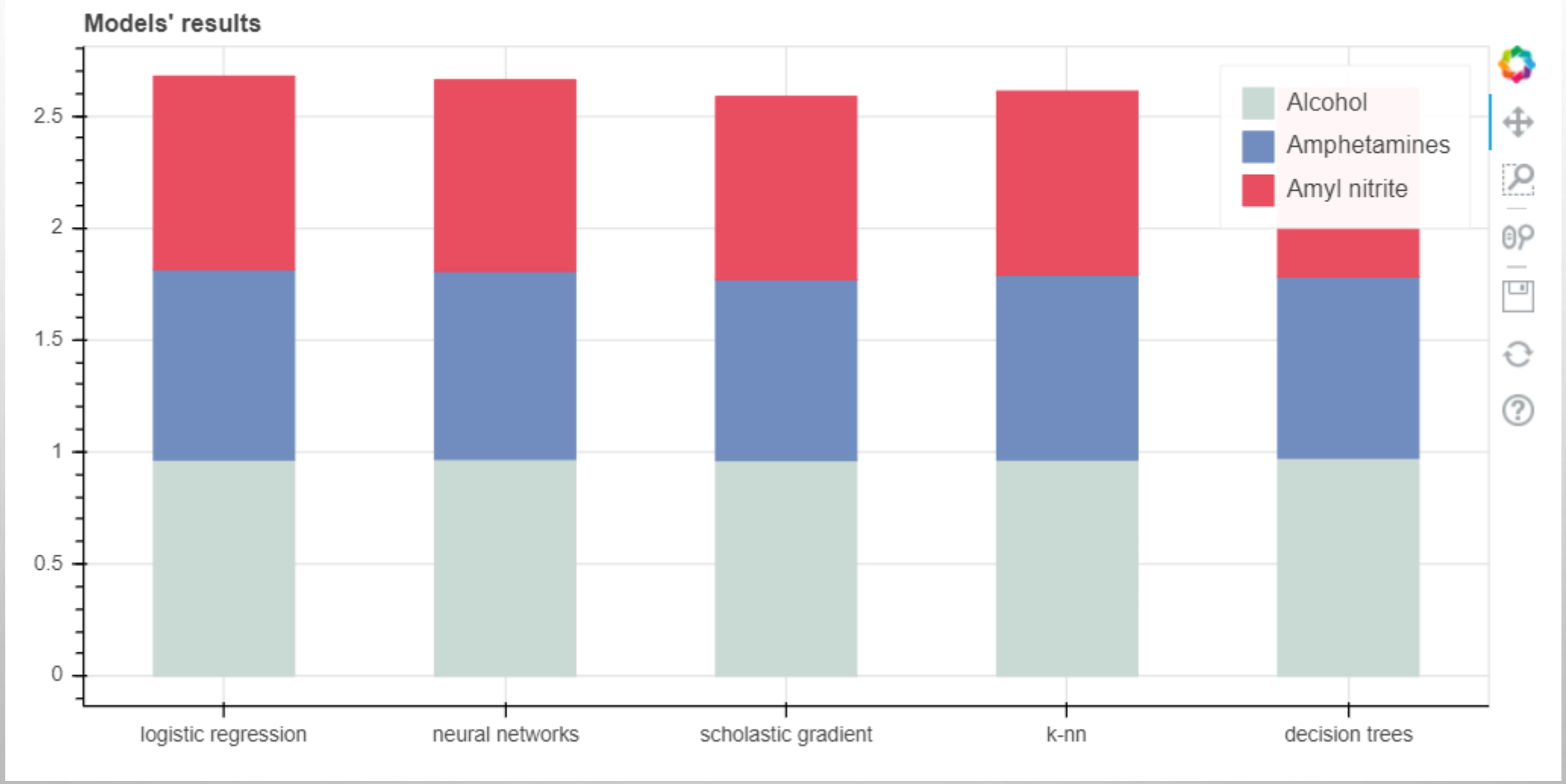
Summary of the results of the models



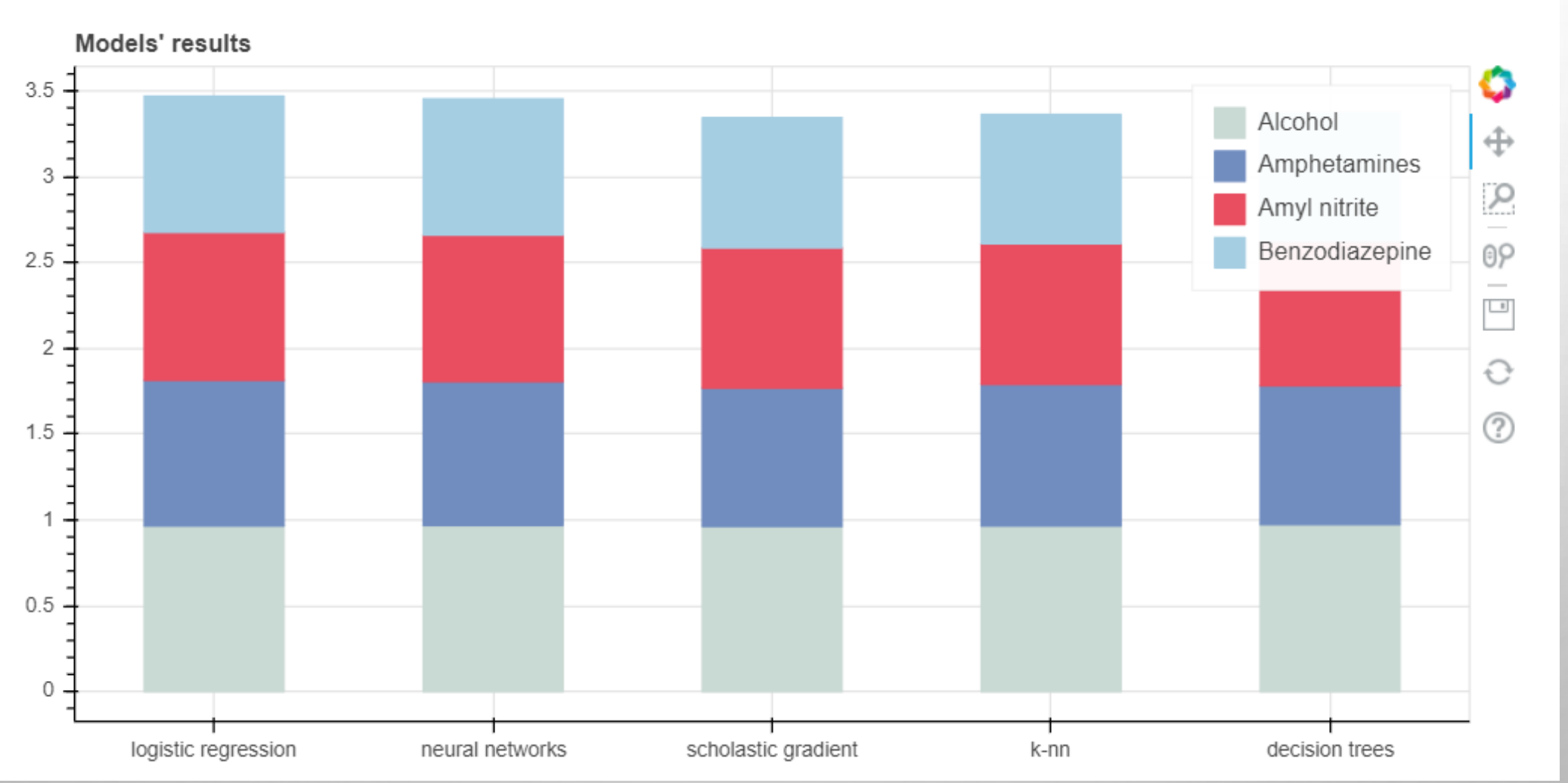
Amphetamines consumption



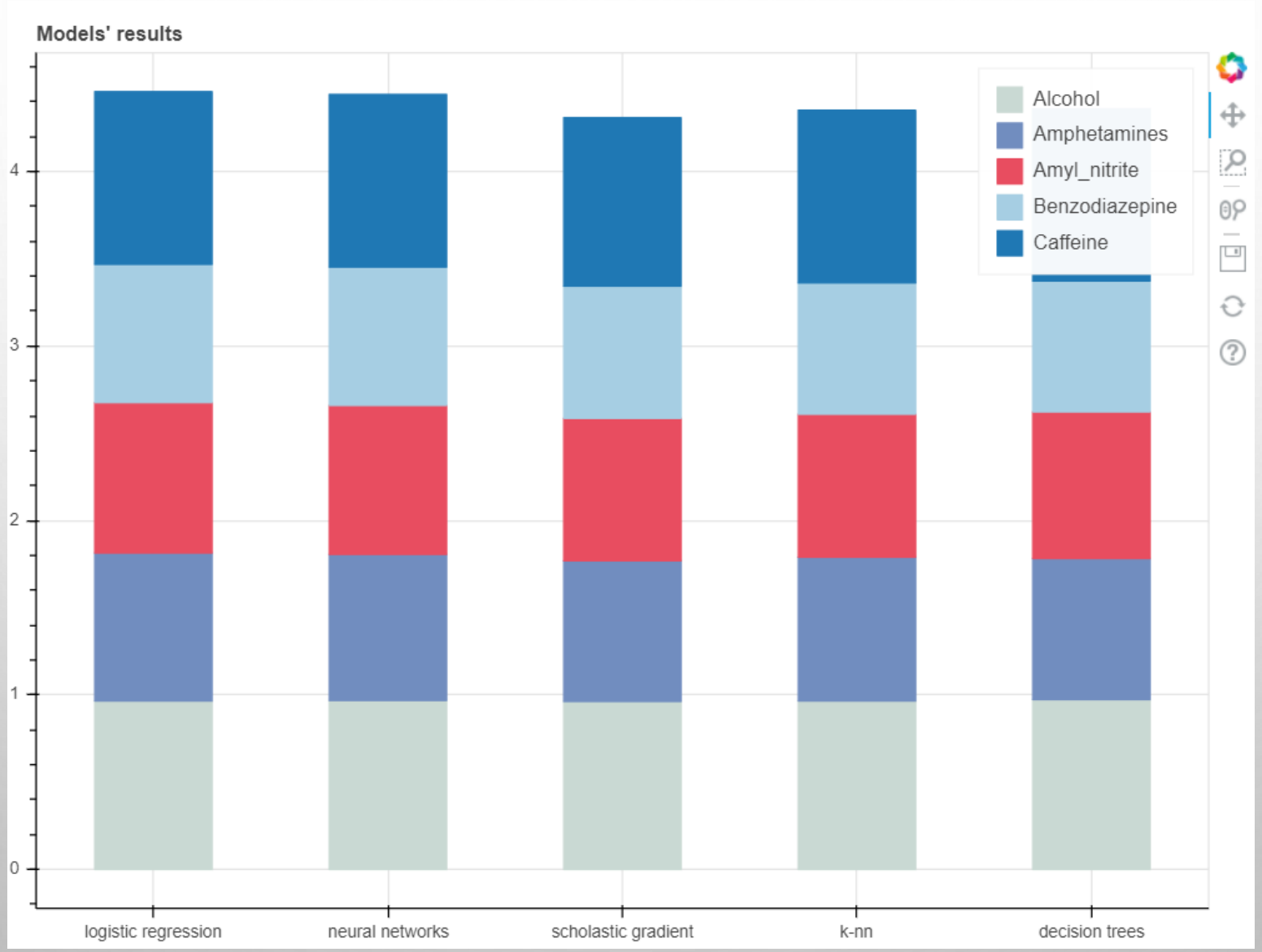
Amyl nitrite consumption



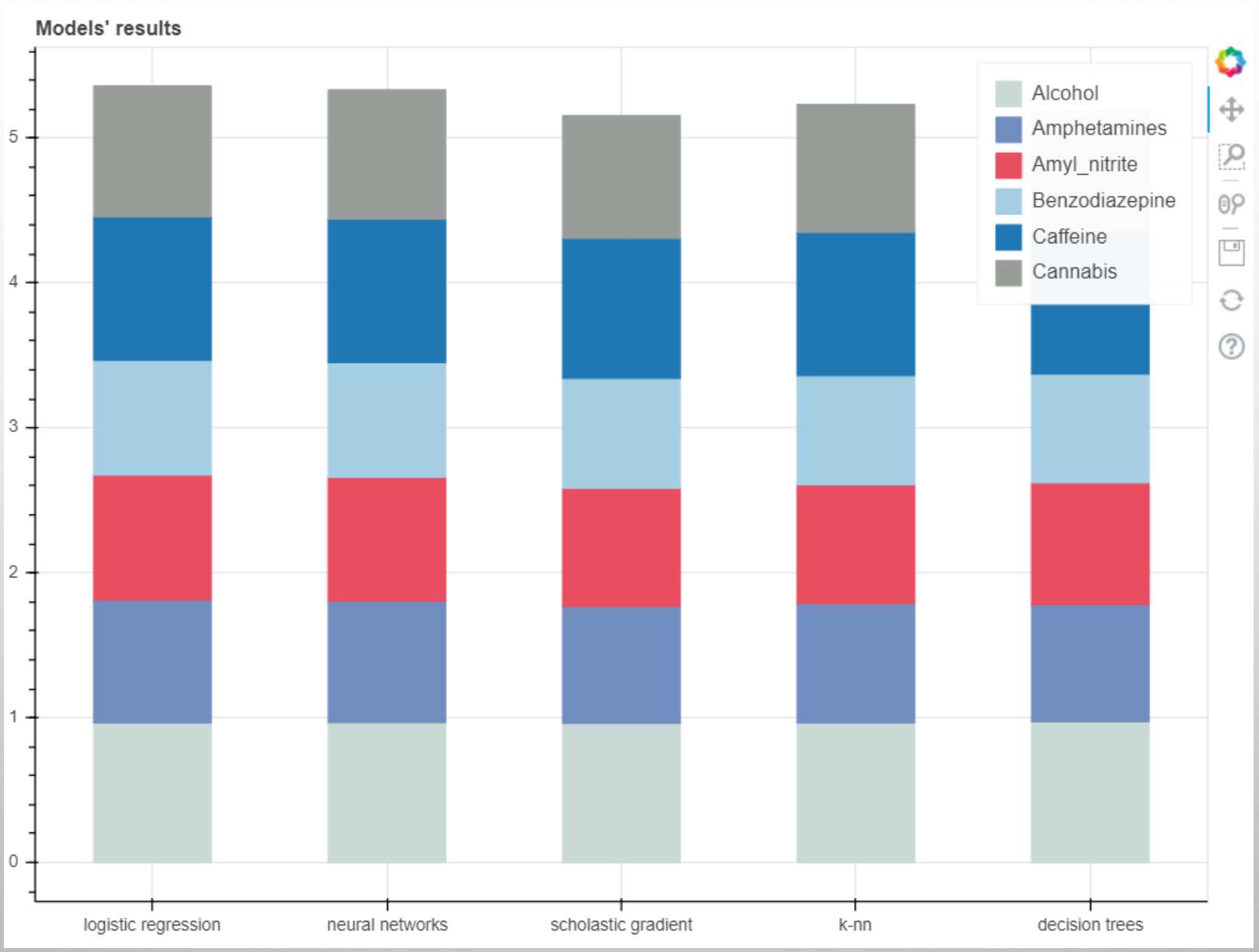
Benzodiazepine consumption



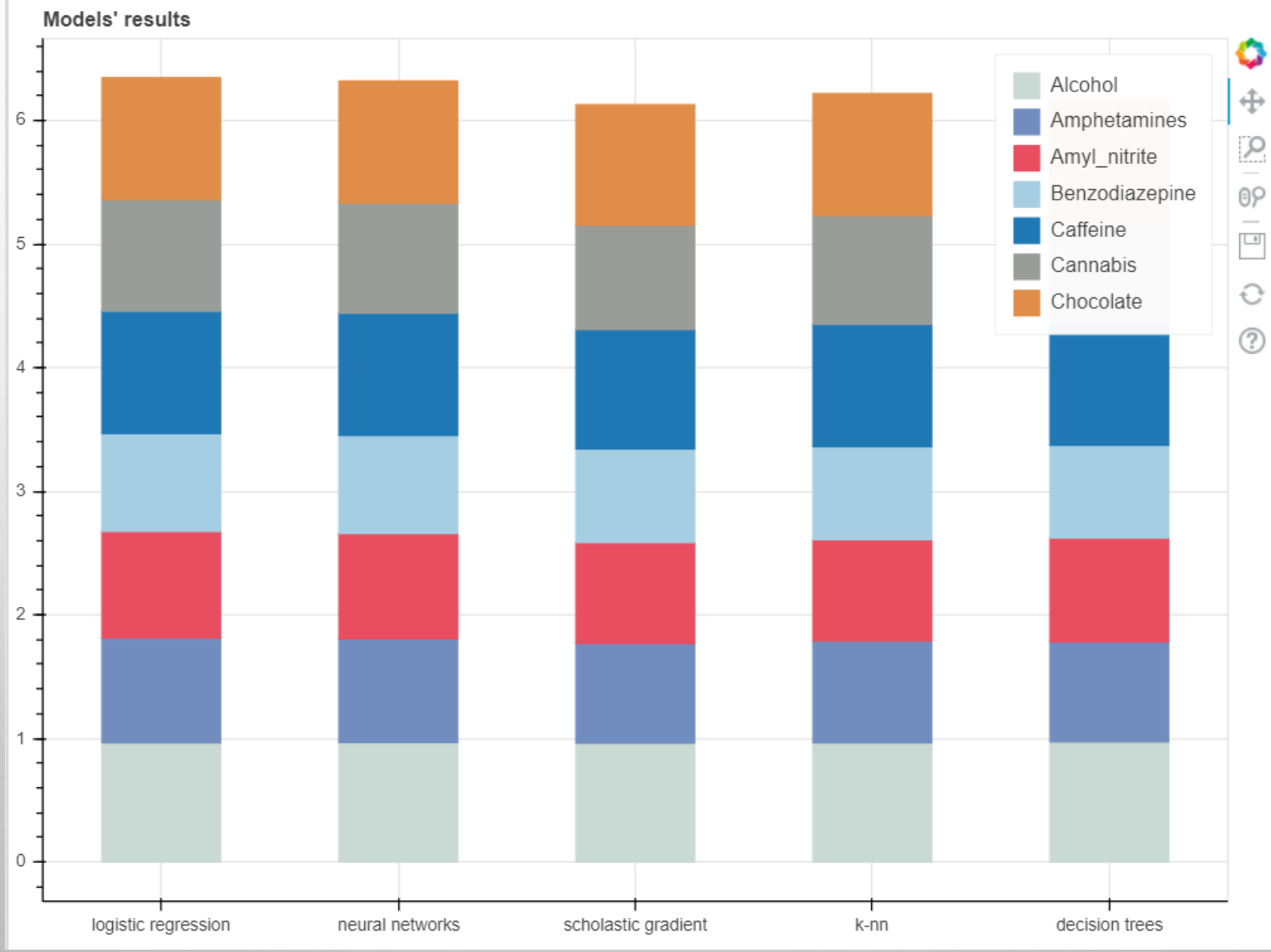
Caffeine consumption



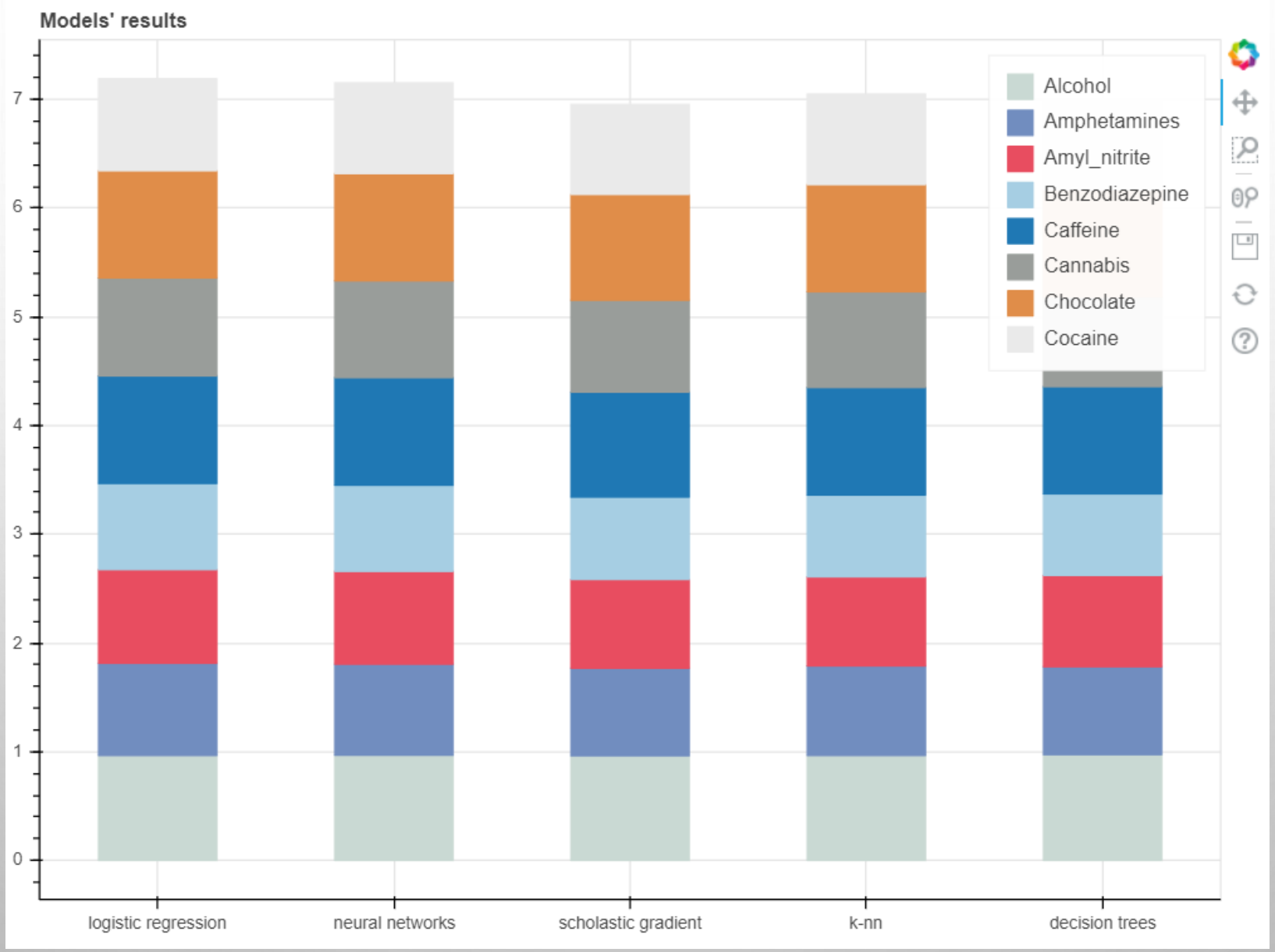
Cannabis consumption



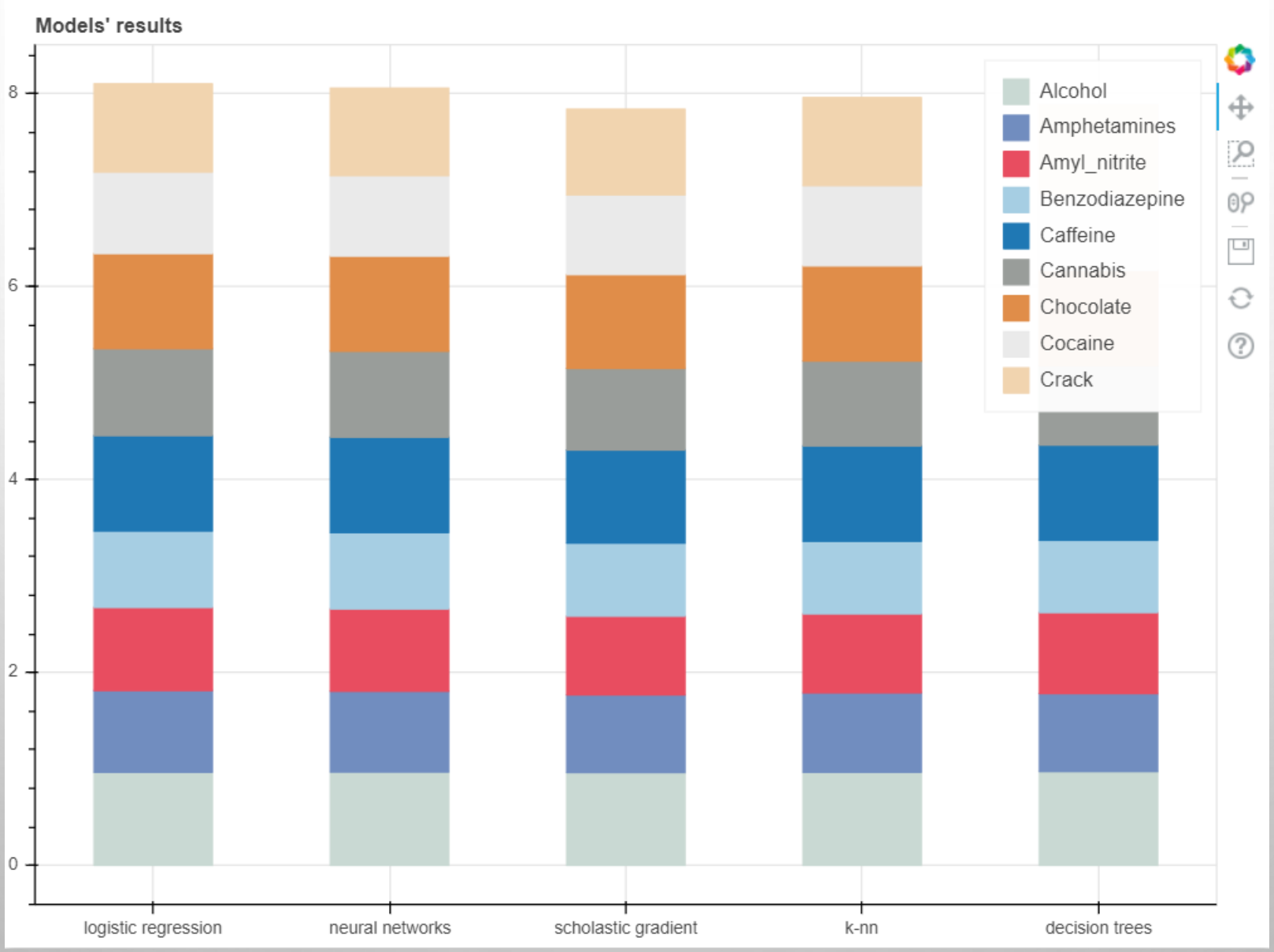
Chocolate consumption



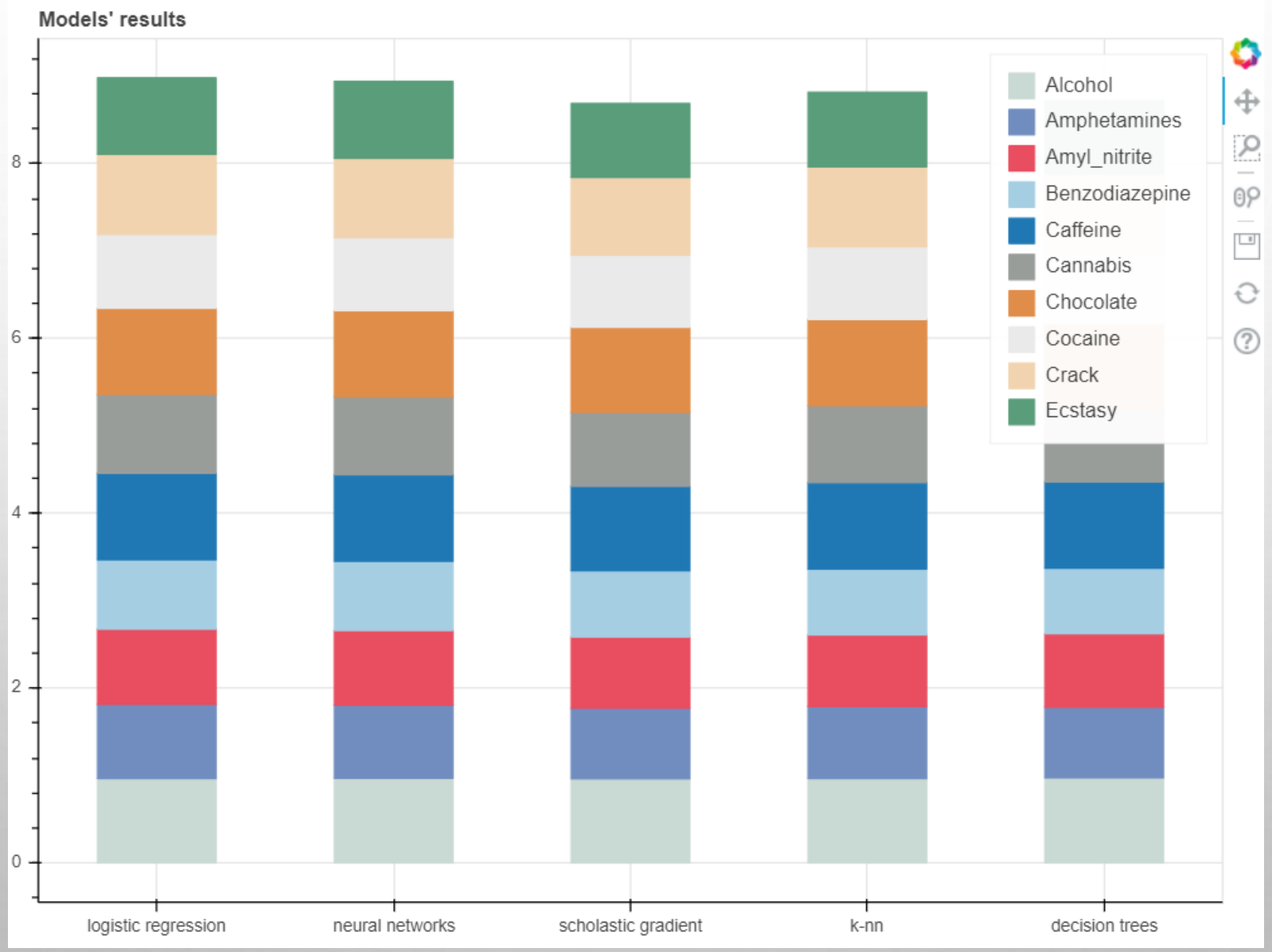
Cocaine consumption



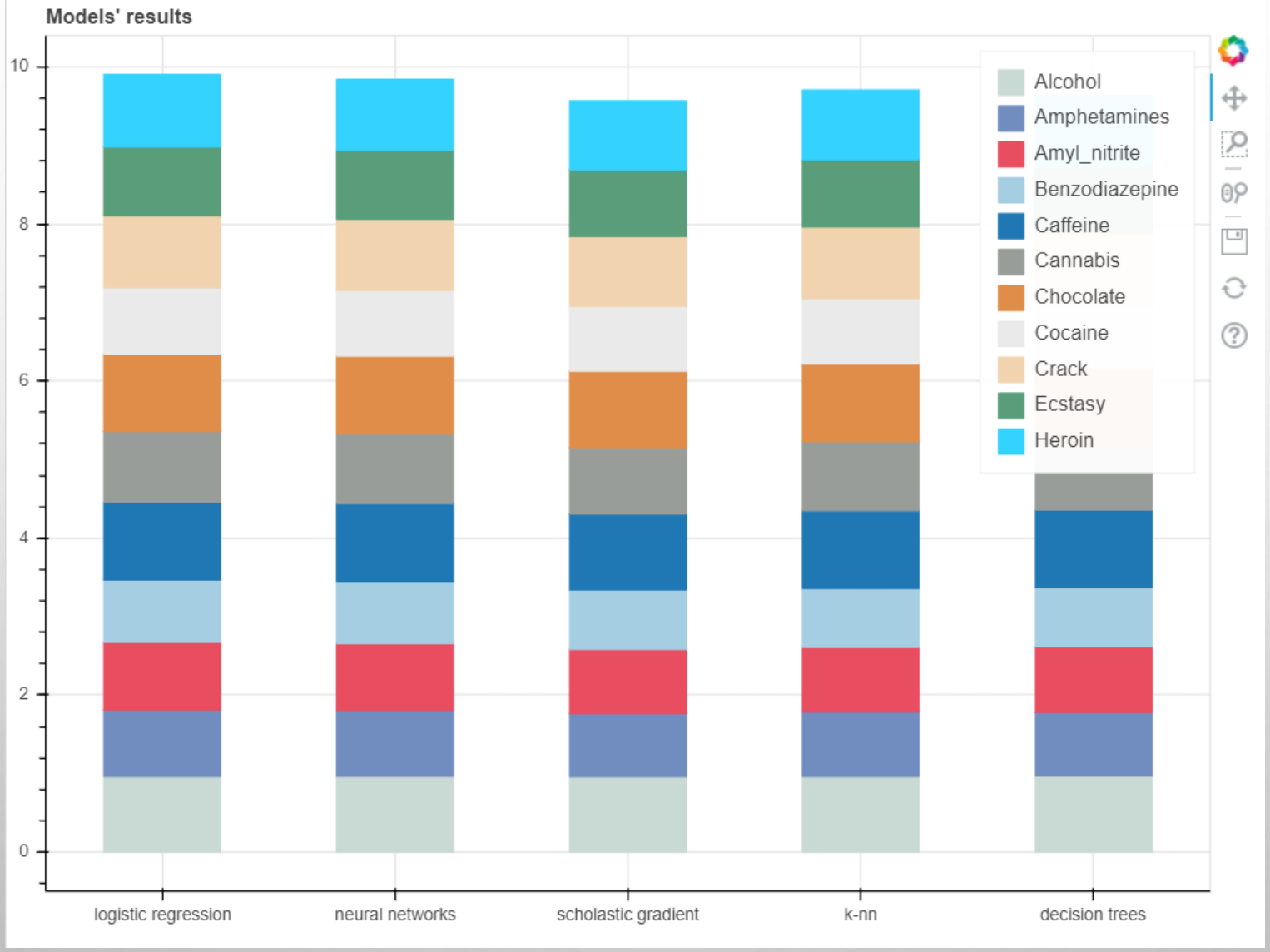
Crack consumption



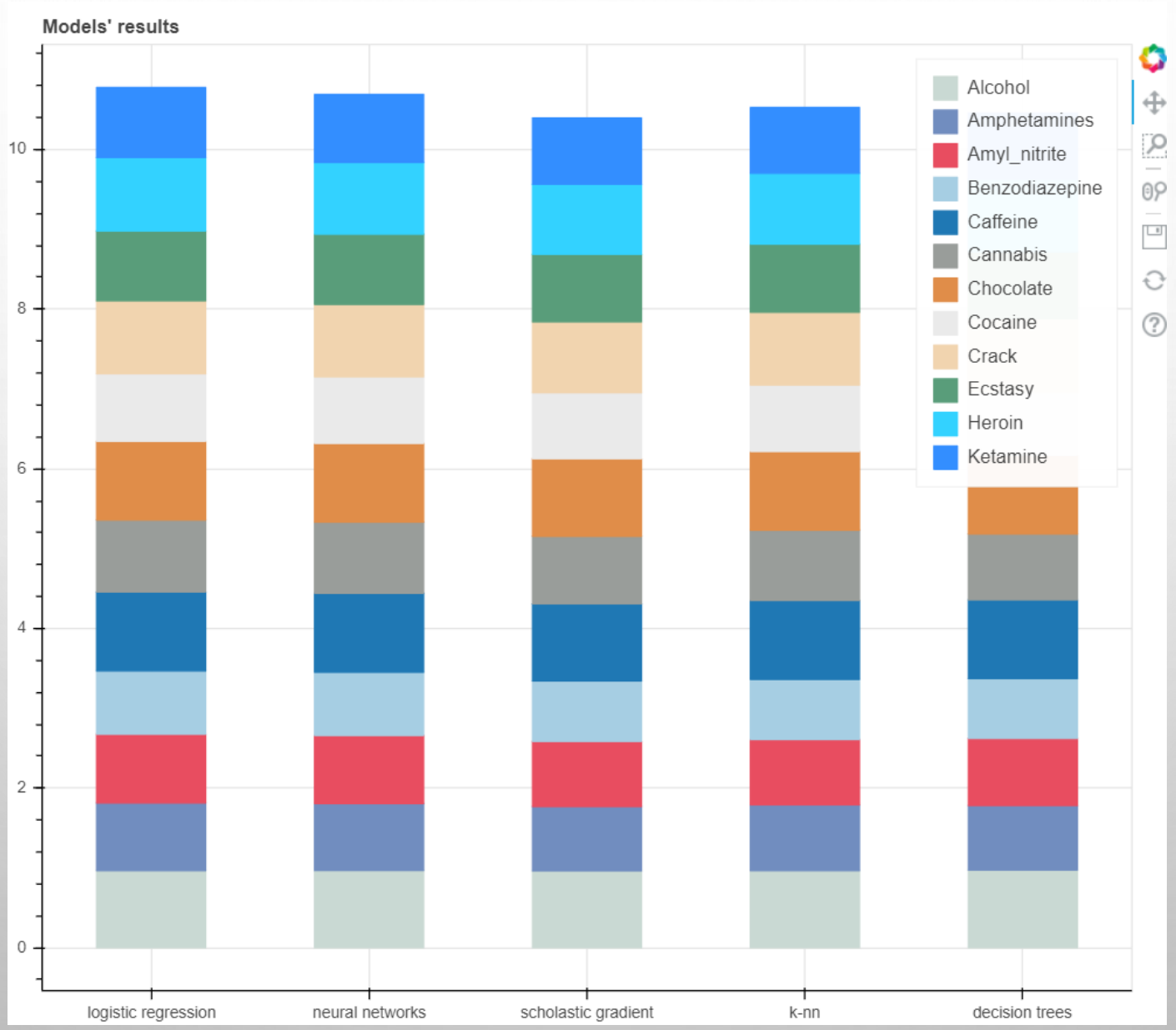
Ecstasy consumption



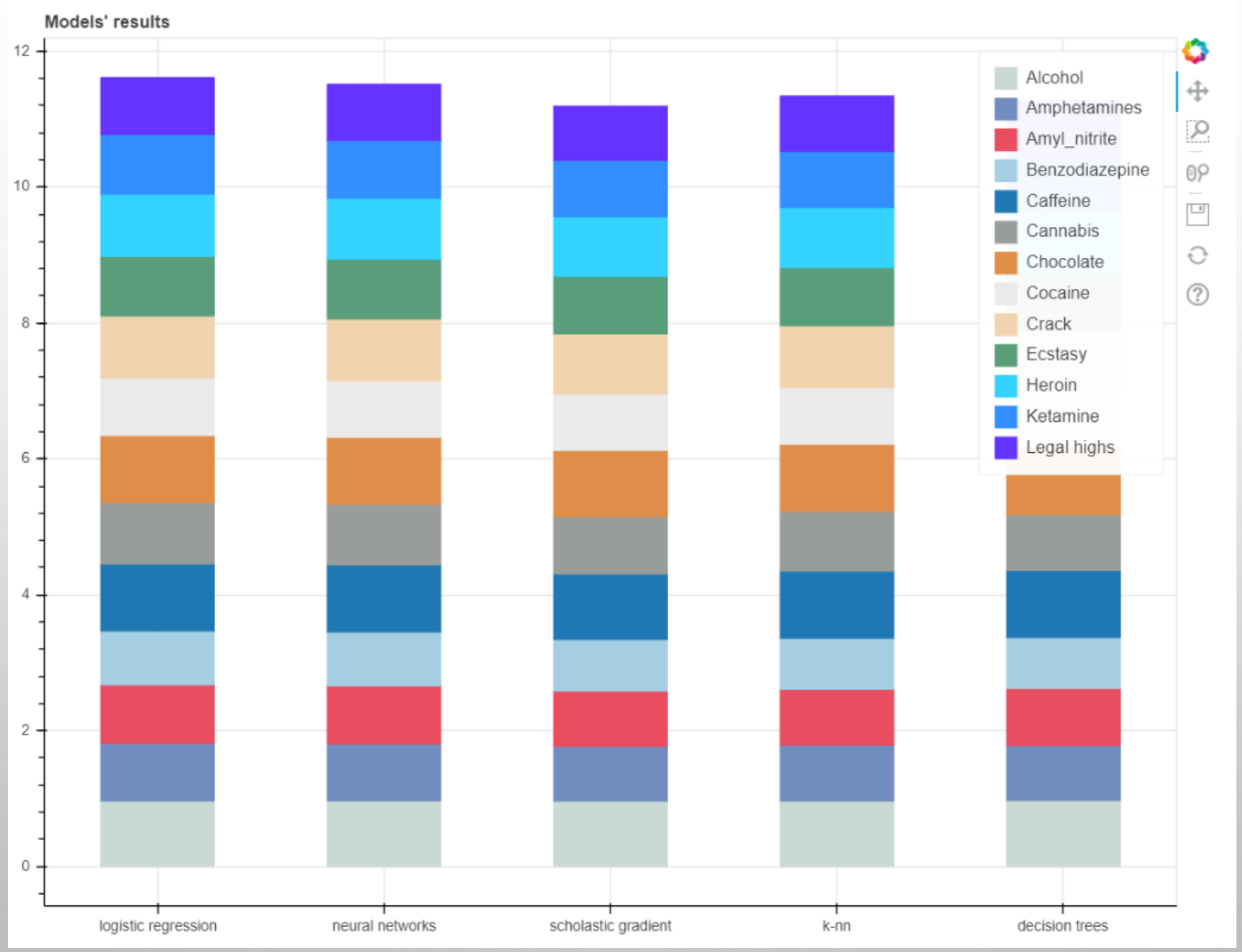
Heroin consumption



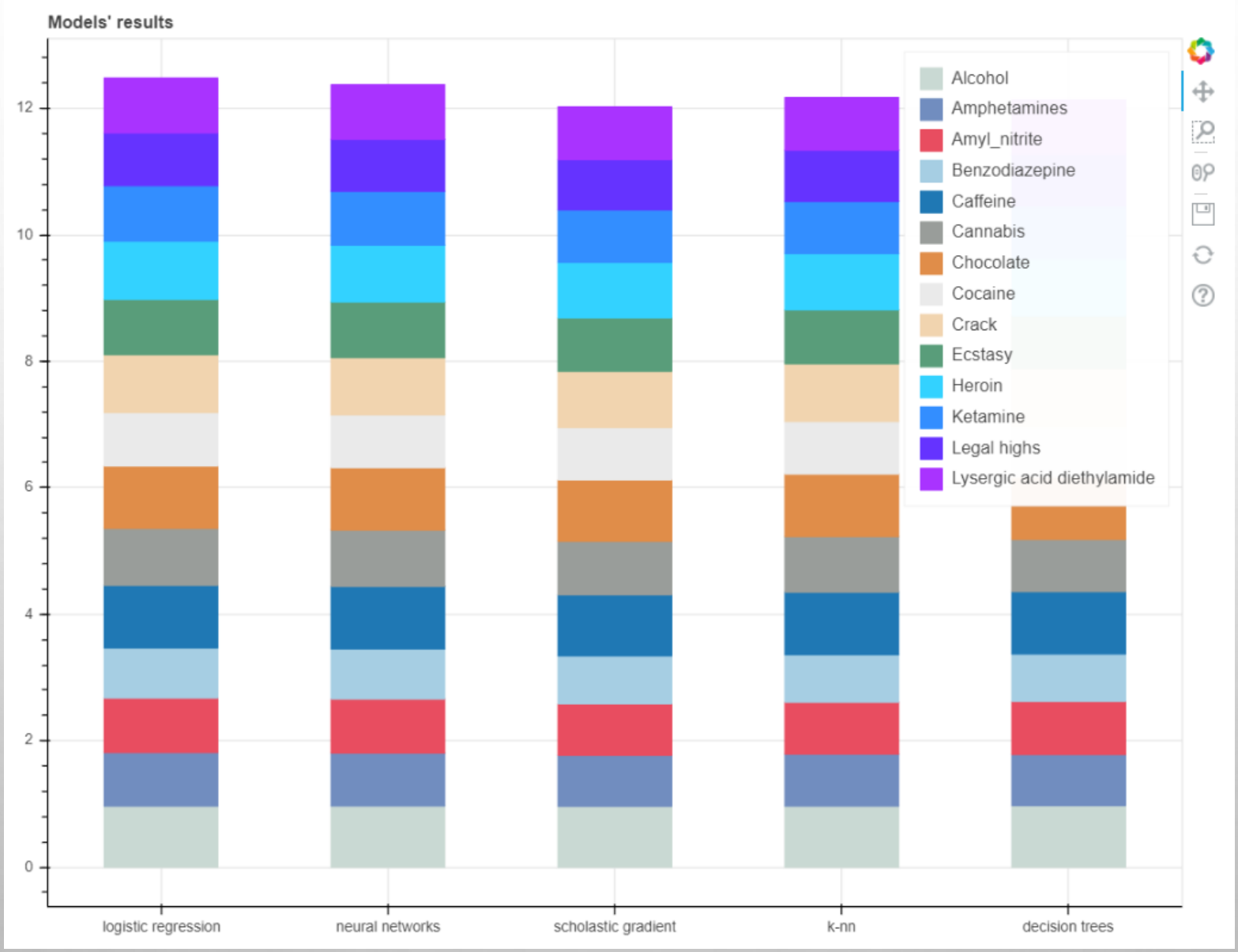
Ketamine consumption



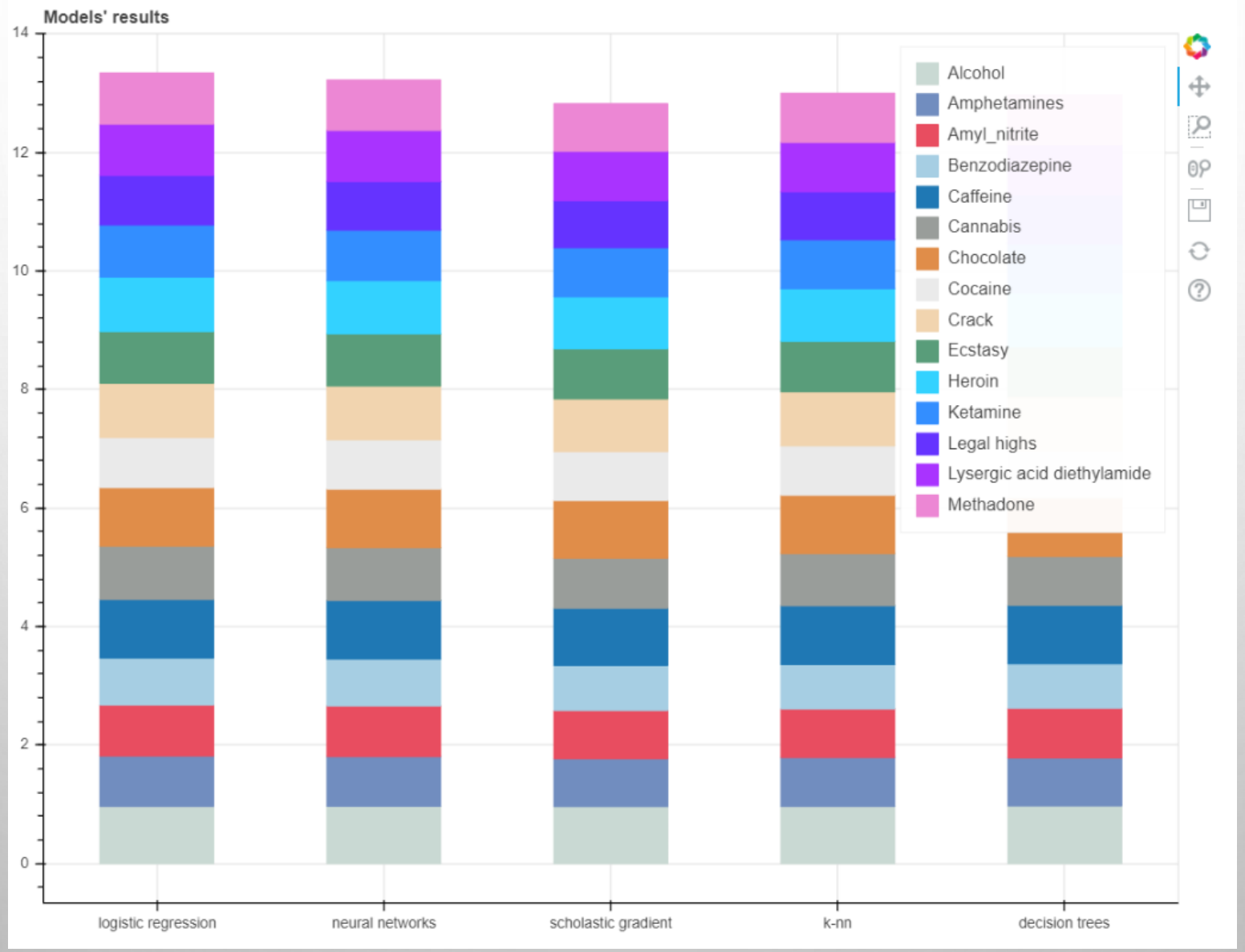
Legal highs consumption



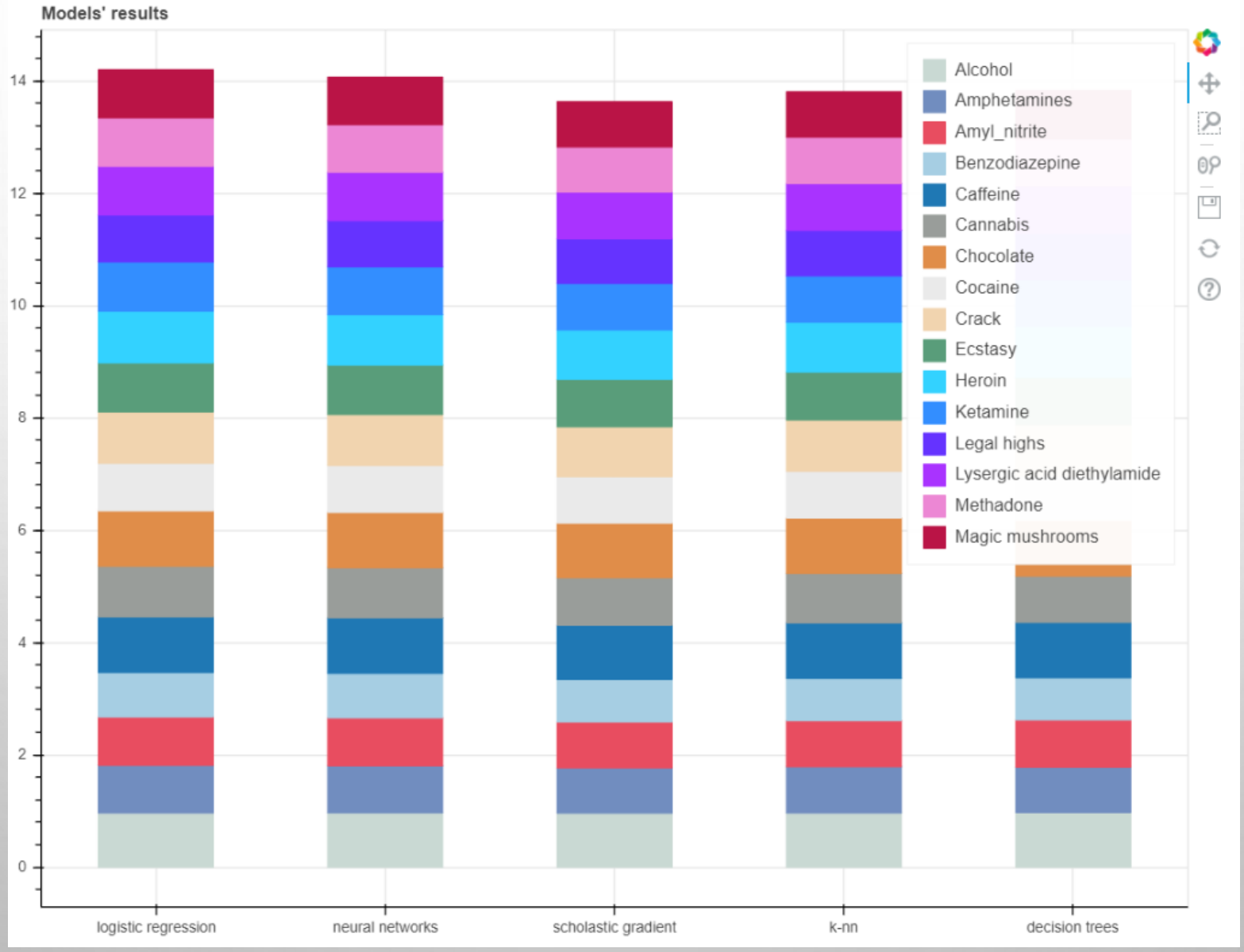
Lysergic acid diethylamide consumption



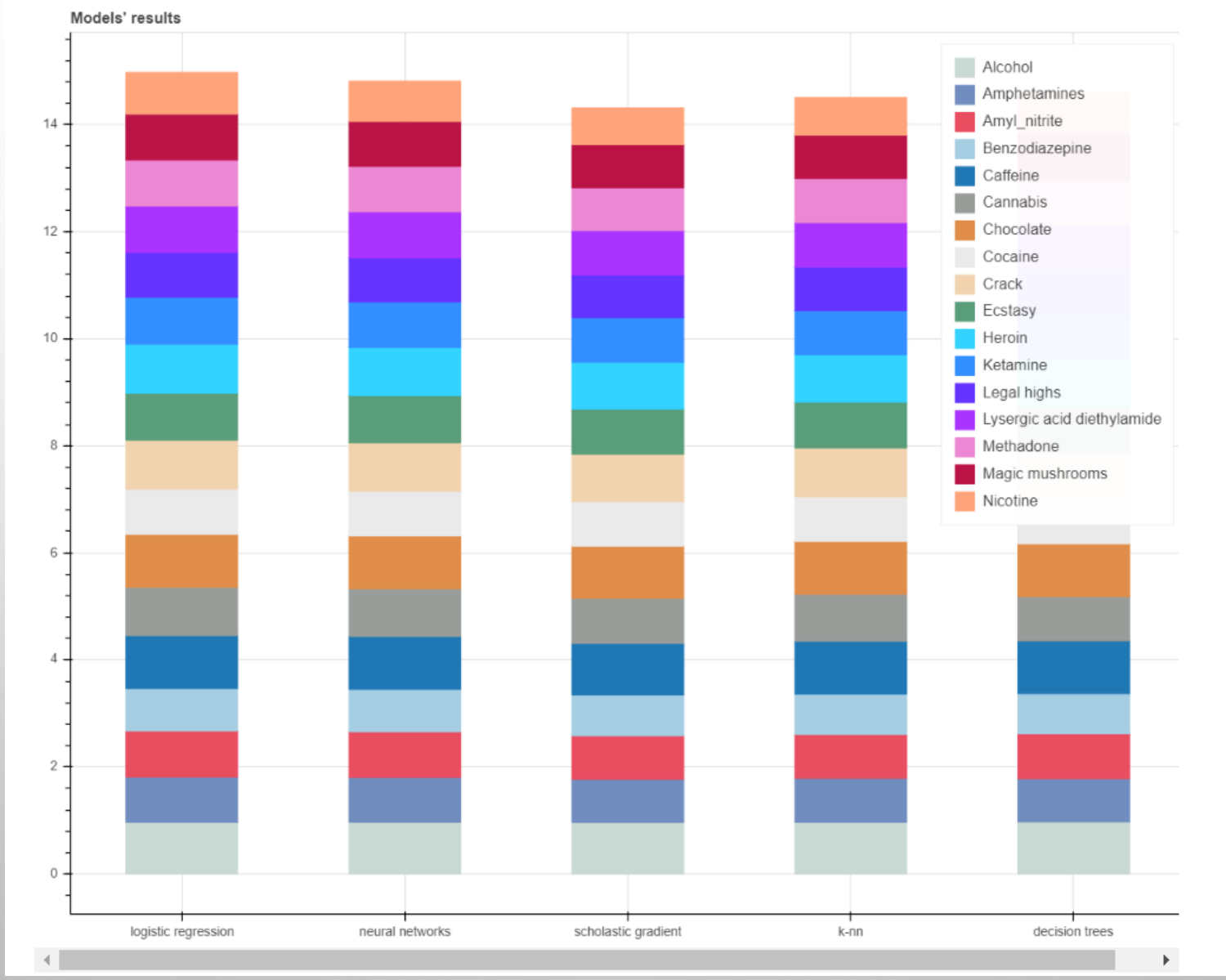
Methadone consumption



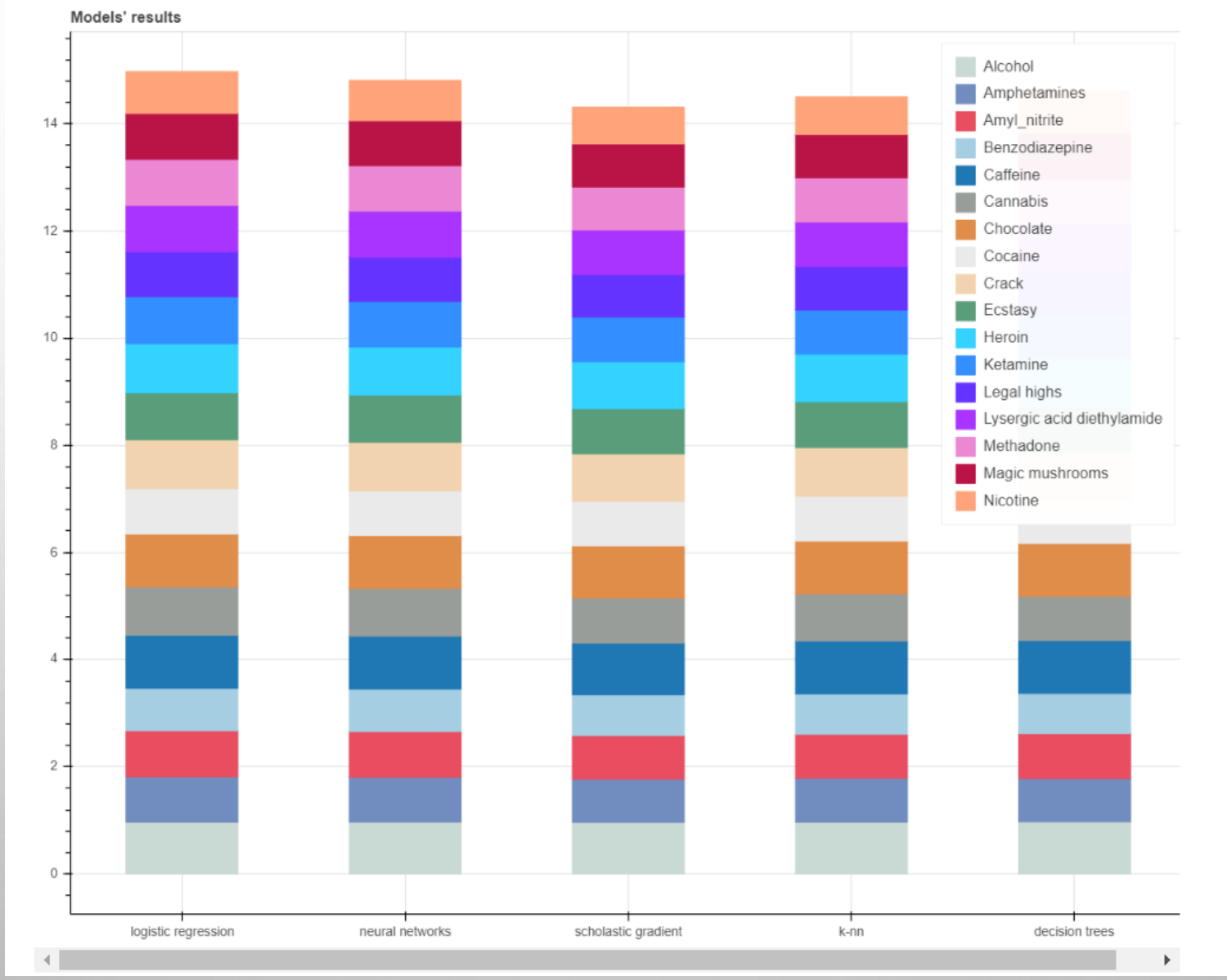
Magic mushrooms consumption



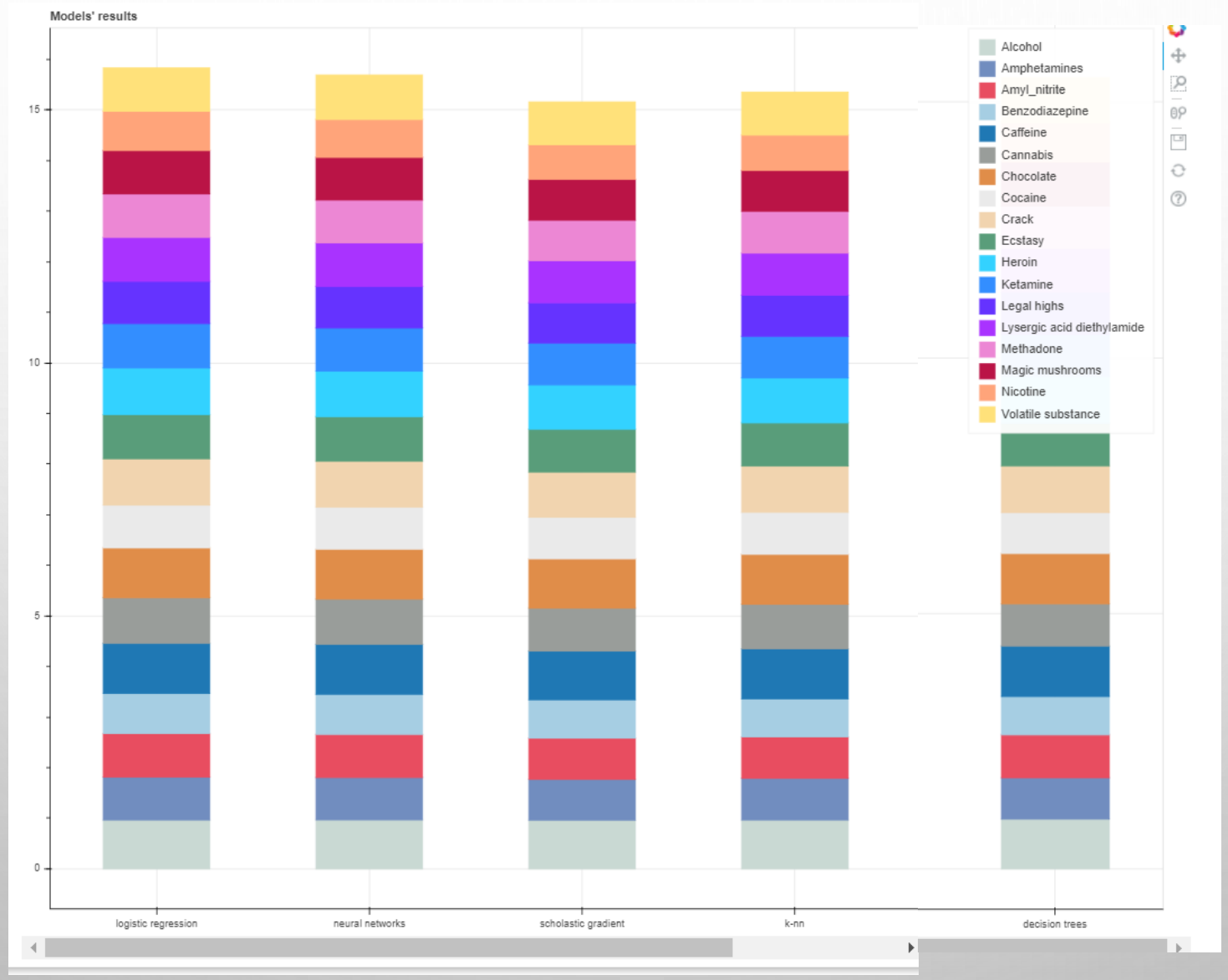
Nicotine consumption



Volatile substance abuse consumption



Nicotine consumption



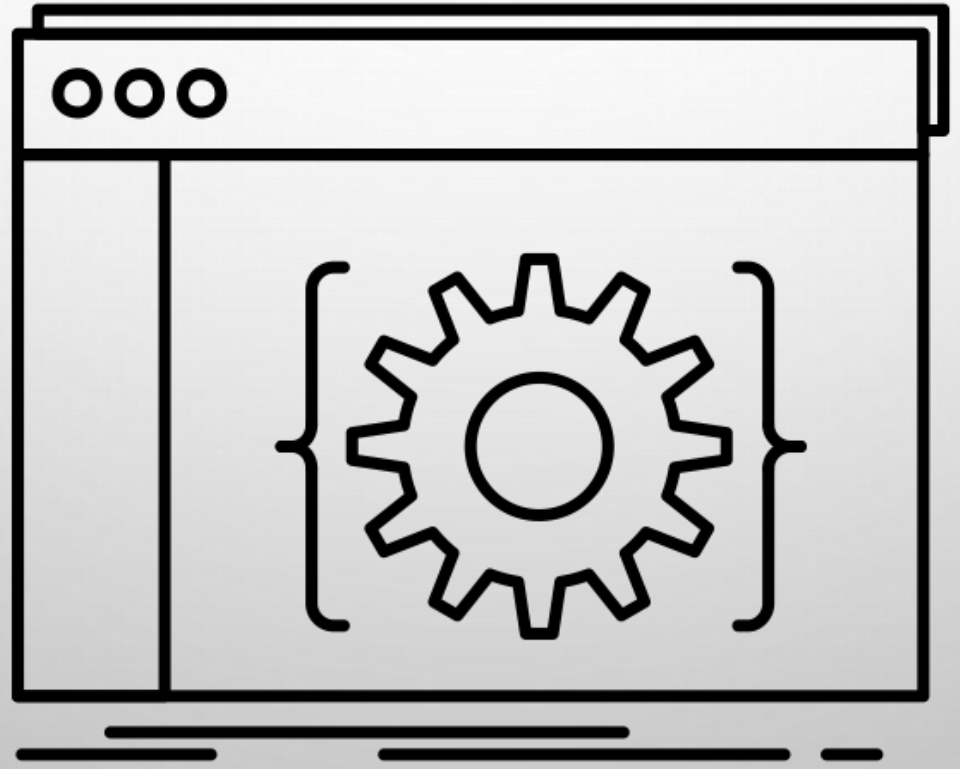
We can observe that if we had to use **one model** to predict the chances of becoming a drug addict for all the drugs, **the logistic regression model** is the most efficient, with **the neural networks model**.

Then come the decision trees model, the k-nn model, and finally the scholastic gradient model. We can still see that the results we obtain from these different models are **really close**.

Plus, the sum of each accuracy of the model are all nearly equal to 16, so we have a mean of **88.9% of accuracy**, which is pretty good.

Finally, please note that even if I read that a lot of people suppressed the chocolate and the coffee since they were pretty common stimulants, I decided to keep them since I wanted to know if the ethnicity or the country could have an influence on the consumption.

TRANSFORMATION OF THE MODEL TO AN API



The flask api

- Take the name of the drug as an argument
- Load the most efficient model
- Take the data cleaned
- Store the model predictions on the data
- Create a pickle that contains the predictions