Dataset Analysis - Drug Consumption

Python for Data Analysis – DIA4

Jesse OHOUENS

Atik MOHAMED MOUKTAR

Table of contents

I. Description of the dataset

II. Dataset Cleaning

III. Data Visualisation

IV. Prediction Models

V. API Flask

https://github.com/Atik14/DrugConsumption_Analysis

Description of the dataset

- Our dataset contains information about how often a person uses different drugs.
- It contains 1885 tuples that correspond to the different people surveyed.
- 13 personality attributes of the respondents are known:

1. ID	2. Age
3. Gender	4. Education
5. Country	6. Ethnicity
7. Nscore (Neuroticism)	8. Escore (Extraversion)
9. Oscore (Openness to experience)	10. Ascore (Agreeableness)
11. Cscore (Conscientiousness)	12. Impulsive (Impulsiveness)
13. Sensation Seeing	

Description of the dataset

- Everyone was asked about their last use of 19 different legal and illegal drugs.

1. Alcohol	2. Amphetamines
3. Amyl nitrite	4. Benzos
5. Caffeine	6. Cannabis
7. Chocolate	8. Cocaine
9. Crack	10. Ecstasy
11. Heroin	12. Ketamine
13. Legal highs	14. LSD
15. Methadone	16. Mushrooms
17. Nicotine	18. Fictitious Drug Semeron
19. Volatile Substance Abuse (VSA)	

Description of the dataset

- They had a choice of 6 options for answers
- For the rest of the study, it is important to specify that we take these answers in a binary way, all those from CL4 onwards are considered as drug users and the others are not. It is a rather strict sharing, but it was a choice in order to have a real idea of the regular consumers like cigarettes.

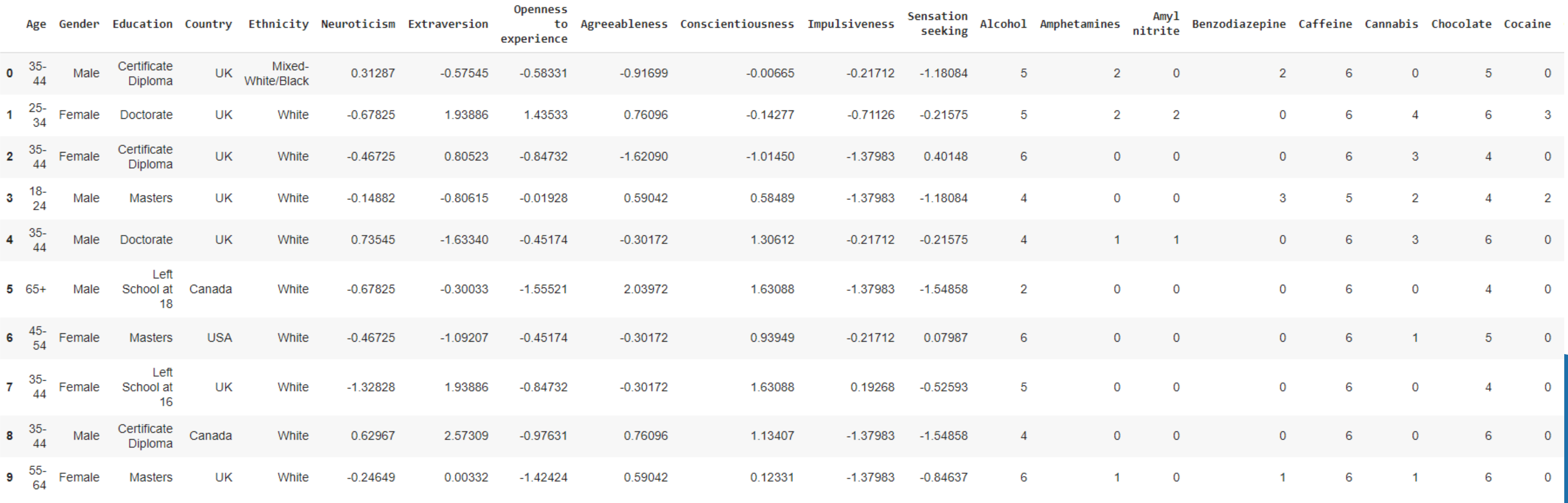
CL0	Never Used
CL1	Used over a Decade Ago
CL2	Used in Last Decade
CL3	Used in Last Year
CL4	Used in Last Month
CL5	Used in Last Week
CL6	Used in Last Day

Dataset Cleaning

- First of all, before starting the analysis, importance must be given to data cleansing. This will ensure the quality of the data to produce reliable and accurate analyses.
- We must be careful with N/A values and remove data that are not relevant to the analysis such as IDs.
- We have also converted the numerical values to nominal values for some classes in order to have more meaningful values.

Dataset Cleaning

- There is a preview of the data after cleaning it.



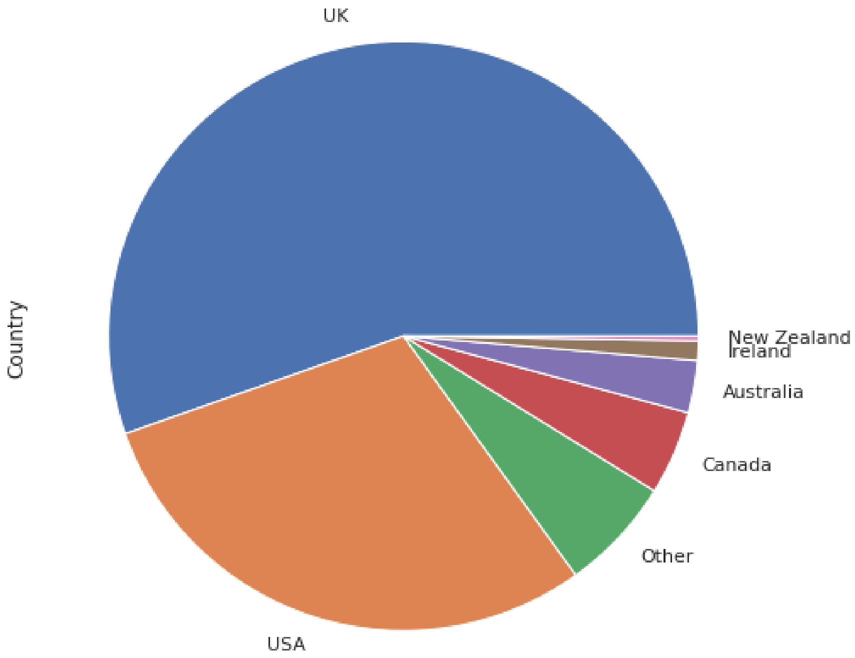
	Age	Gender	Education	Country	Ethnicity	Neuroticism	Extraversion	Openness to experience	Agreeableness	Conscientiousness	Impulsiveness	Sensation seeking	Alcohol	Amphetamines	Amyl nitrite	Benzodiazepine	Caffeine	Cannabis	Chocolate	Cocaine
0	35-44	Male	Certificate Diploma	UK	Mixed-White/Black	0.31287	-0.57545	-0.58331	-0.91699	-0.00665	-0.21712	-1.18084	5	2	0	2	6	0	5	0
1	25-34	Female	Doctorate	UK	White	-0.67825	1.93886	1.43533	0.76096	-0.14277	-0.71126	-0.21575	5	2	2	0	6	4	6	3
2	35-44	Female	Certificate Diploma	UK	White	-0.46725	0.80523	-0.84732	-1.62090	-1.01450	-1.37983	0.40148	6	0	0	0	6	3	4	0
3	18-24	Male	Masters	UK	White	-0.14882	-0.80615	-0.01928	0.59042	0.58489	-1.37983	-1.18084	4	0	0	3	5	2	4	2
4	35-44	Male	Doctorate	UK	White	0.73545	-1.63340	-0.45174	-0.30172	1.30612	-0.21712	-0.21575	4	1	1	0	6	3	6	0
5	65+	Male	Left School at 18	Canada	White	-0.67825	-0.30033	-1.55521	2.03972	1.63088	-1.37983	-1.54858	2	0	0	0	6	0	4	0
6	45-54	Female	Masters	USA	White	-0.46725	-1.09207	-0.45174	-0.30172	0.93949	-0.21712	0.07987	6	0	0	0	6	1	5	0
7	35-44	Female	Left School at 16	UK	White	-1.32828	1.93886	-0.84732	-0.30172	1.63088	0.19268	-0.52593	5	0	0	0	6	0	4	0
8	35-44	Male	Certificate Diploma	Canada	White	0.62967	2.57309	-0.97631	0.76096	1.13407	-1.37983	-1.54858	4	0	0	0	6	0	6	0
9	55-64	Female	Masters	UK	White	-0.24649	0.00332	-1.42424	0.59042	0.12331	-1.37983	-0.84637	6	1	0	1	6	1	6	0

Data Visualisation

- This dataset contains a lot of information about individuals.
- We will thus represent these data in the form of graphs in order to better observe the distribution of the various types of individuals.

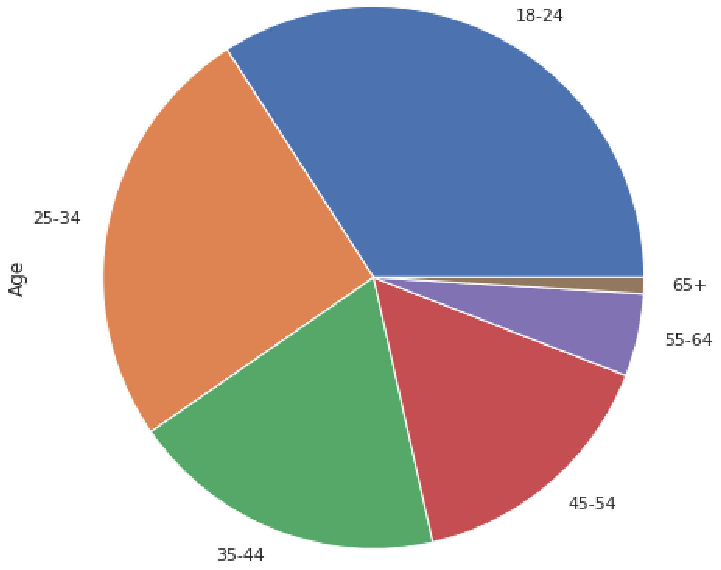
Data Visualisation

- Here is a graph of the distribution of individuals according to their country.
- We can see that they come mainly from the United Kingdom and the United States.
- We have principally here the population of only 6 countries, so it does not represent the world population.



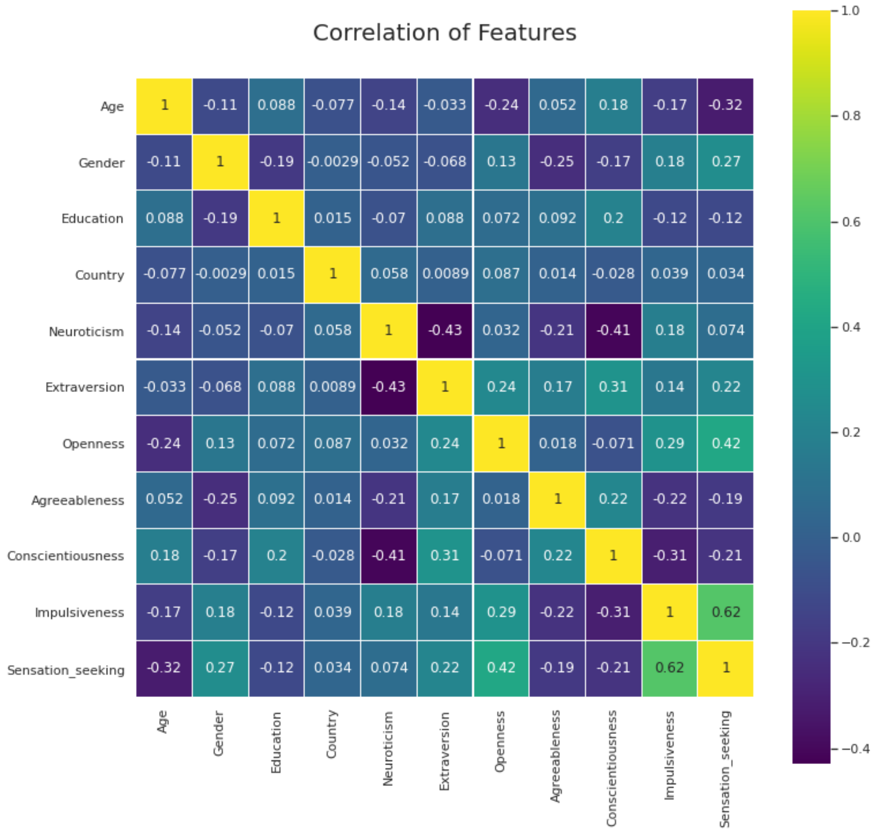
Data Visualisation

- Here is a graph of the distribution of individuals according to their age group.
- Most of the individuals interviewed are rather young, they are between 18 and 34 years old.



Data Visualisation

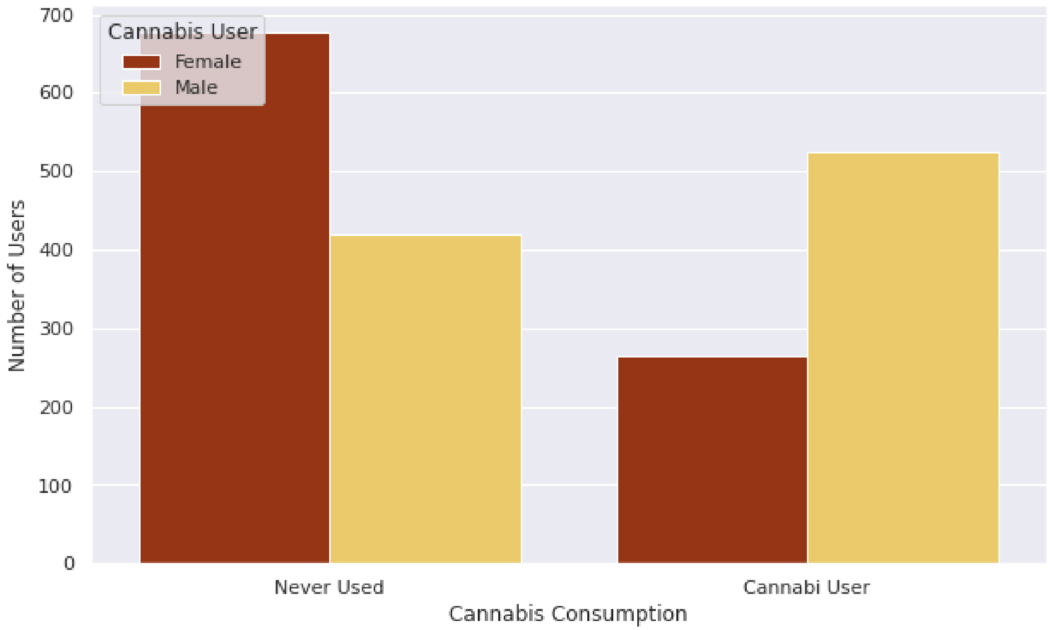
- Here is a graph of correlations between the different personality attributes of an individual.
- In particular, we can see that neuroticism is negatively correlated with extraversion and conscientiousness.


Correlation of Features

	Age	Gender	Education	Country	Neuroticism	Extraversion	Openness	Agreeableness	Conscientiousness	Impulsiveness	Sensation_seeking
Age	1	-0.11	0.088	-0.077	-0.14	-0.033	-0.24	0.052	0.18	-0.17	-0.32
Gender	-0.11	1	-0.19	-0.0029	-0.052	-0.068	0.13	-0.25	-0.17	0.18	0.27
Education	0.088	-0.19	1	0.015	-0.07	0.088	0.072	0.092	0.2	-0.12	-0.12
Country	-0.077	-0.0029	0.015	1	0.058	0.0089	0.087	0.014	-0.028	0.039	0.034
Neuroticism	-0.14	-0.052	-0.07	0.058	1	-0.43	0.032	-0.21	-0.41	0.18	0.074
Extraversion	-0.033	-0.068	0.088	0.0089	-0.43	1	0.24	0.17	0.31	0.14	0.22
Openness	-0.24	0.13	0.072	0.087	0.032	0.24	1	0.018	-0.071	0.29	0.42
Agreeableness	0.052	-0.25	0.092	0.014	-0.21	0.17	0.018	1	0.22	-0.22	-0.19
Conscientiousness	0.18	-0.17	0.2	-0.028	-0.41	0.31	-0.071	0.22	1	-0.31	-0.21
Impulsiveness	-0.17	0.18	-0.12	0.039	0.18	0.14	0.29	-0.22	-0.31	1	0.62
Sensation_seeking	-0.32	0.27	-0.12	0.034	0.074	0.22	0.42	-0.19	-0.21	0.62	1

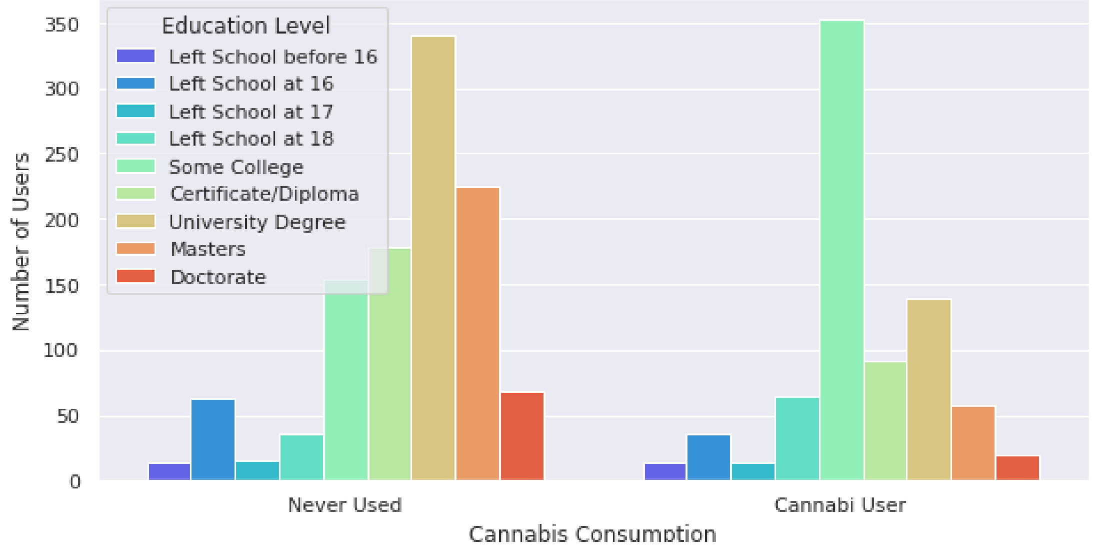
Data Visualisation

- Here is a graph of cannabis use by gender.
- It can be seen that men use more cannabis than women.
- We considered an individual to be a cannabis user if he or she has used cannabis at least this month.



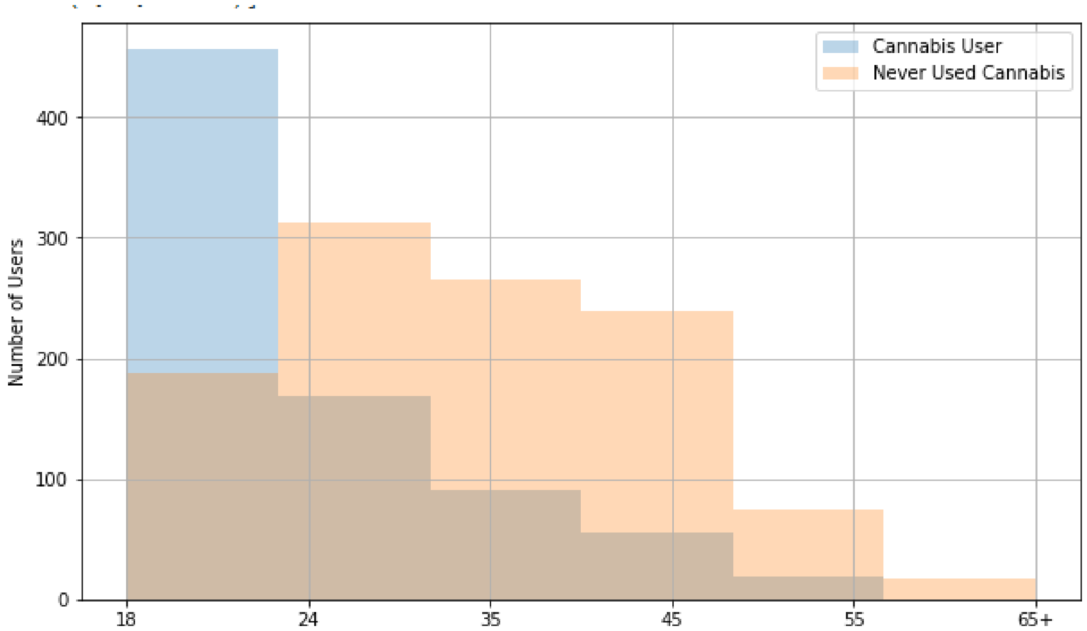
Data Visualisation

- Here is a graph of cannabis use according to an individual's level of education.
- It can be observed that cannabis users have mainly completed a few years of college and that non-users have mainly university degrees and masters degrees.



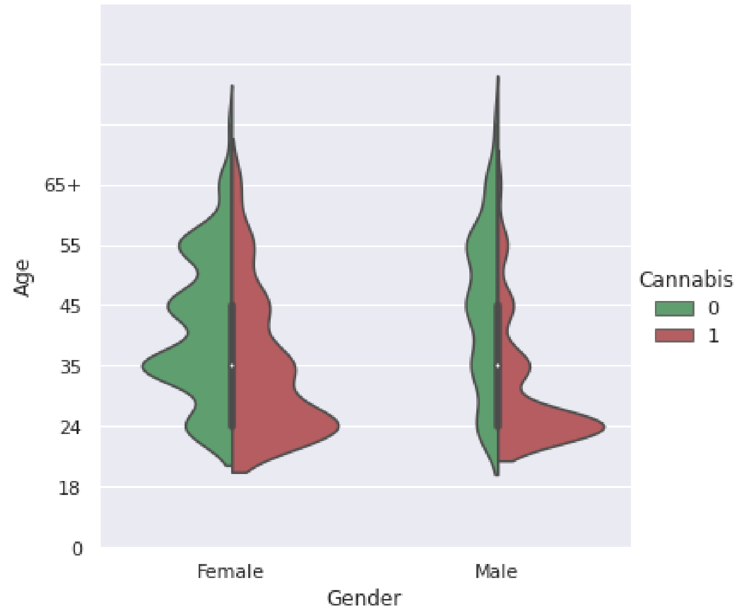
Data Visualisation

- Here is a graph of cannabis use or never used cannabis according to different age groups.
- It can be observed that cannabis users have mainly in the 18-24 age groups. It can already be concluded that the regular use of cannabis comes mainly from relatively young people.



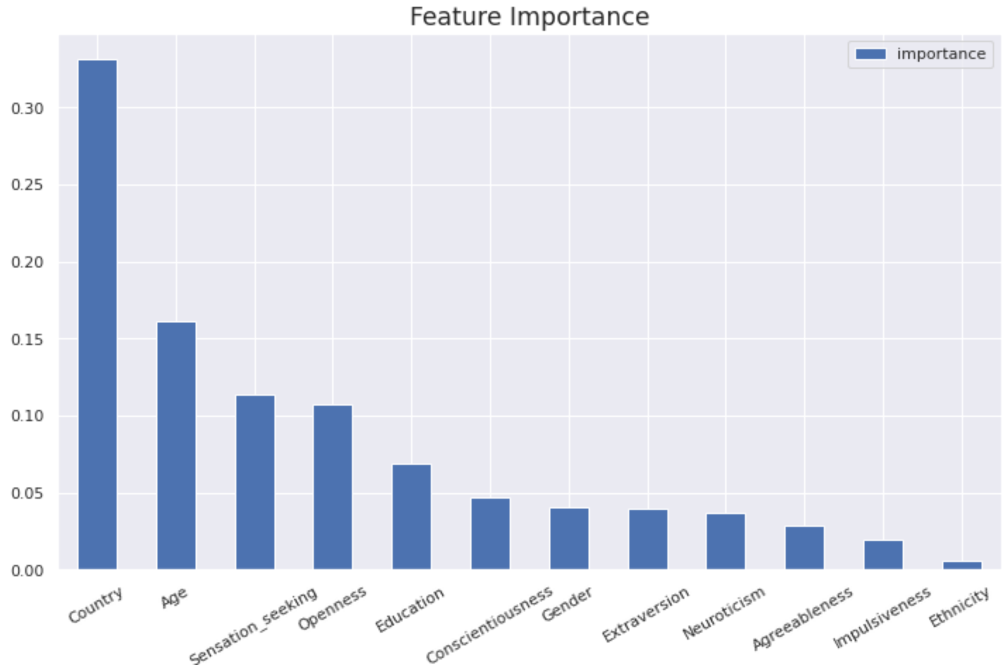
Data Visualisation

- Here is a graph of cannabis use in relation to age in combination with gender.
- We can observe that the previous observations are confirmed, and this allows a more global visualization of the target individual that we are gradually starting to create : "The young male" (0 = Not Use Cannabis, 1 = Use Cannabis)



Data Visualisation

- Here is a graph of the importance of different features.
- We can observe that the country is an essential element in the consumption of cannabis, which is quite logical, given that some states allow it. Somewhat more generally the demographic elements are relatively more important than the other elements.



Prediction Models

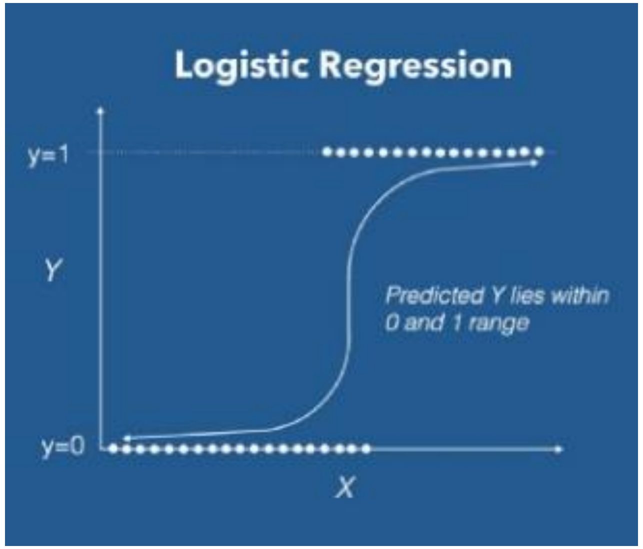
- From this dataset, there were many possible predictions according to the different drugs.
- In our case, we will focus more specifically on **cannabis use** because it is a drug that is becoming more and more legalized in different countries, such as Canada recently in 2018.
- We have therefore conducted an analysis based on an individual's attributes and the frequency with which he or she uses different drugs in order to determine if he or she is a cannabis user.

Prediction Models

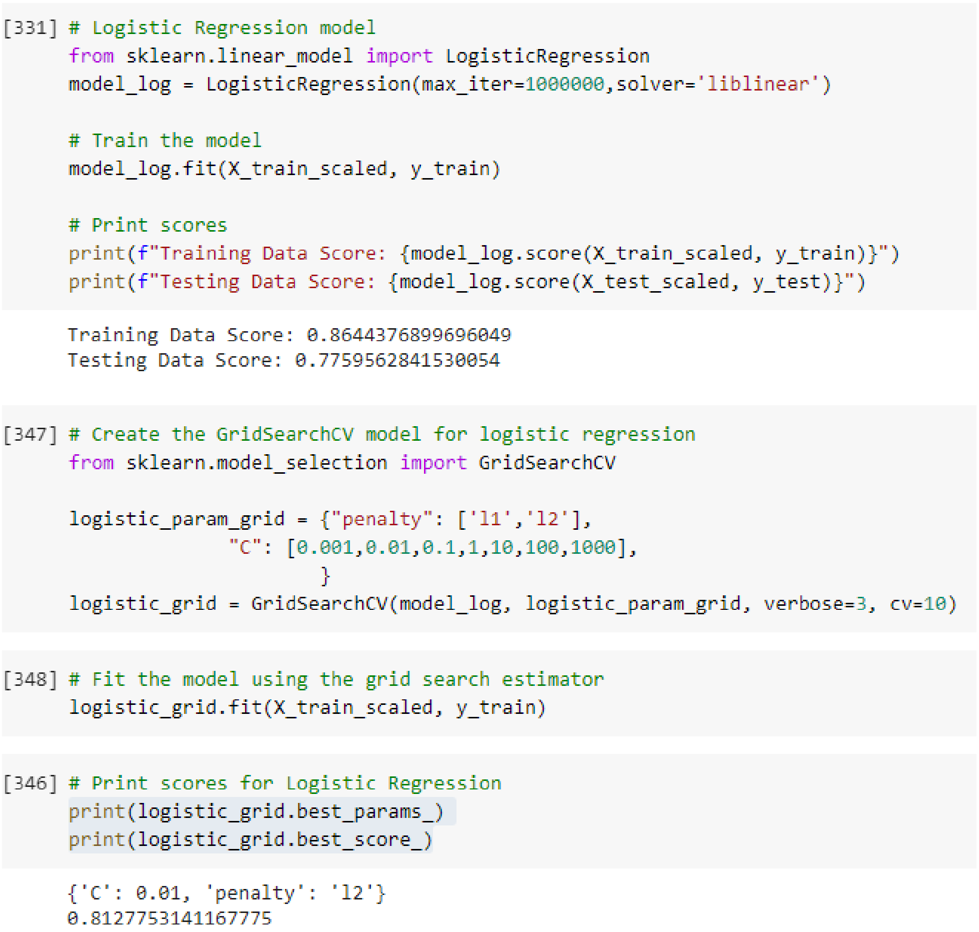
- Our Team’s goal was to find the best machine learning model to predict cannabis users when looking at features such personality traits and demographics.
- Find and apply the machine learning model with the most accurate prediction the data set and predict the potential risk of cannabis use
- The target prediction is to determine if the selected features influence the use of cannabis among our data respondents.

Prediction Models

- Among the different possible prediction model possibilities, we chose to transform the dataset into a binary classification problem comprising 2 classes: **Cannabis User** and **Cannabis Non-User**.
- To solve it, we immediately thought of using a logistic regression model.
- **Logistic Regression** is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.



Prediction Models



```
[331] # Logistic Regression model
      from sklearn.linear_model import LogisticRegression
      model_log = LogisticRegression(max_iter=1000000,solver='liblinear')

      # Train the model
      model_log.fit(X_train_scaled, y_train)

      # Print scores
      print(f"Training Data Score: {model_log.score(X_train_scaled, y_train)}")
      print(f"Testing Data Score: {model_log.score(X_test_scaled, y_test)}")

      Training Data Score: 0.8644376899696049
      Testing Data Score: 0.7759562841530054
```

```
[347] # Create the GridSearchCV model for logistic regression
      from sklearn.model_selection import GridSearchCV

      logistic_param_grid = {"penalty": ['l1','l2'],
                    "C": [0.001,0.01,0.1,1,10,100,1000],
                          }
      logistic_grid = GridSearchCV(model_log, logistic_param_grid, verbose=3, cv=10)
```

```
[348] # Fit the model using the grid search estimator
      logistic_grid.fit(X_train_scaled, y_train)
```

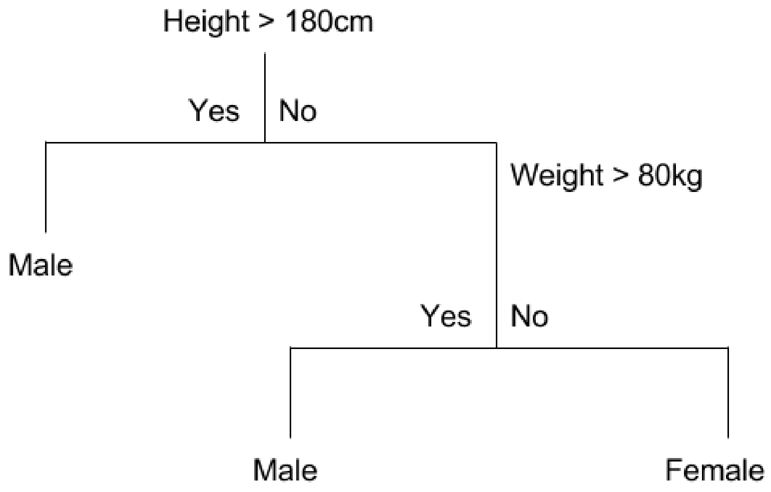
```
[346] # Print scores for Logistic Regression
      print(logistic_grid.best_params_)
      print(logistic_grid.best_score_)

      {'C': 0.01, 'penalty': 'l2'}
      0.8127753141167775
```

Logistic Regression

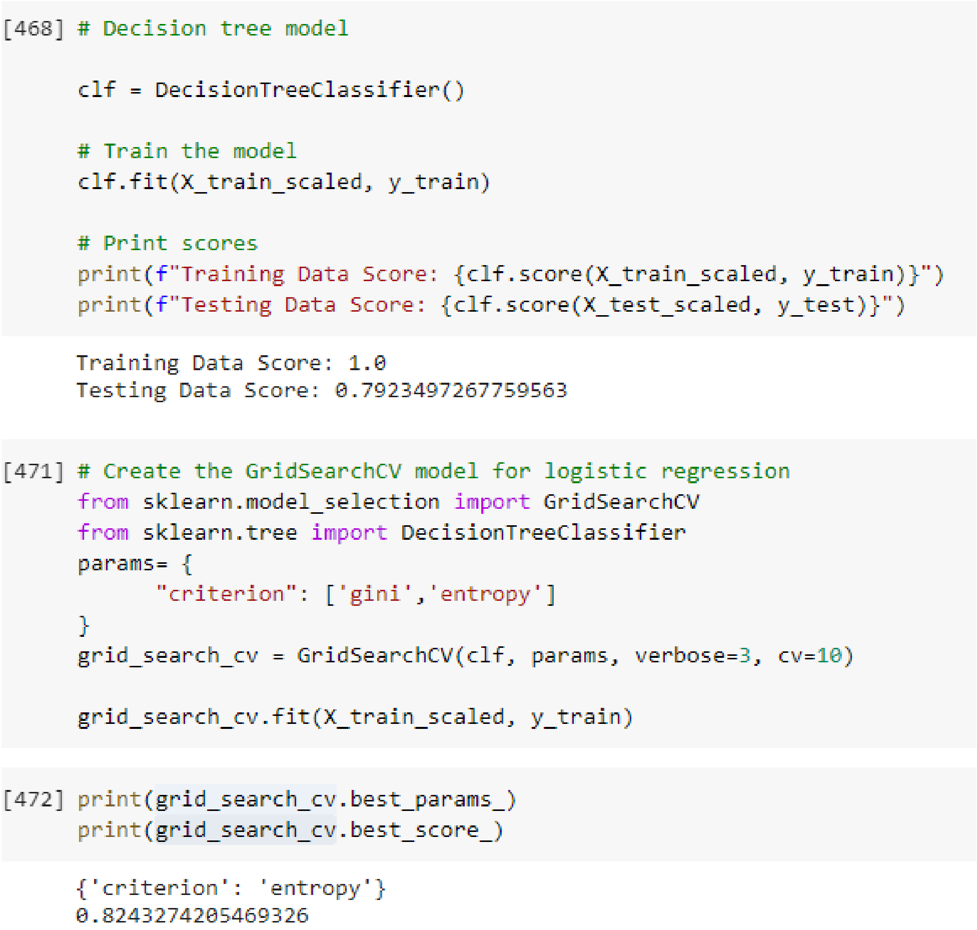
- Accuracy : 0.8127753141167775

Prediction Models

- Then, we also tried a Decision Tree prediction model to compare accuracies.
- A **Decision Tree** is a simple representation for classifying examples. It is a Supervised Machine Learning where the data is continuously split according to a certain parameter.



Prediction Models



```
[468] # Decision tree model

      clf = DecisionTreeClassifier()

      # Train the model
      clf.fit(X_train_scaled, y_train)

      # Print scores
      print(f"Training Data Score: {clf.score(X_train_scaled, y_train)}")
      print(f"Testing Data Score: {clf.score(X_test_scaled, y_test)}")

      Training Data Score: 1.0
      Testing Data Score: 0.7923497267759563
```

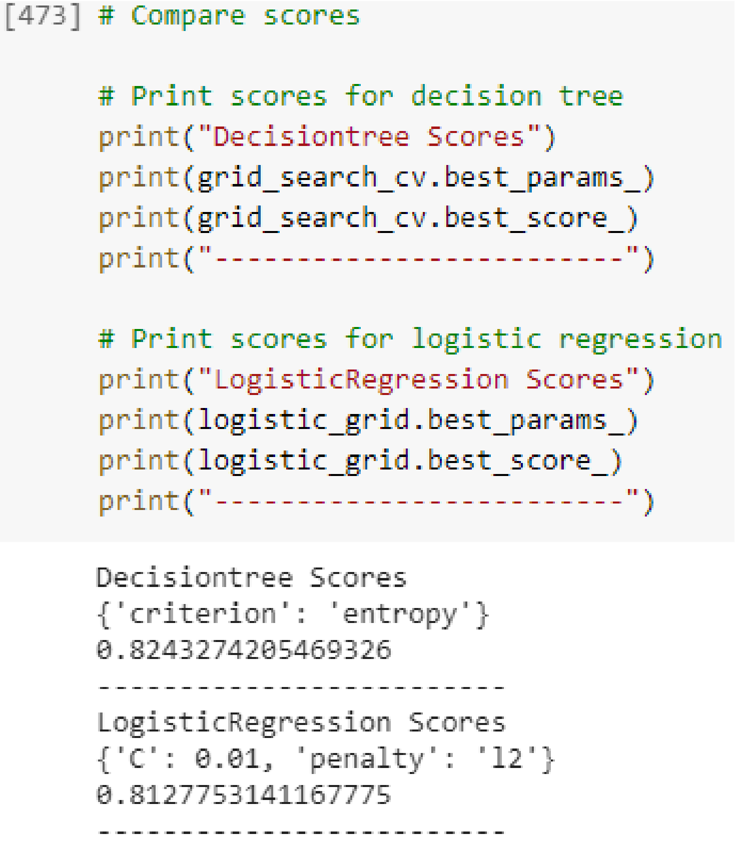
```
[471] # Create the GridSearchCV model for logistic regression
      from sklearn.model_selection import GridSearchCV
      from sklearn.tree import DecisionTreeClassifier
      params= {
            "criterion": ['gini','entropy']
      }
      grid_search_cv = GridSearchCV(clf, params, verbose=3, cv=10)

      grid_search_cv.fit(X_train_scaled, y_train)
```

```
[472] print(grid_search_cv.best_params_)
      print(grid_search_cv.best_score_)

      {'criterion': 'entropy'}
      0.8243274205469326
```

Decision tree

- Accuracy : 0.8243274205469326

Model Results comparison



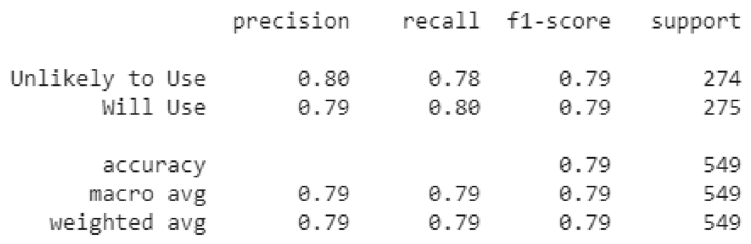
```
[473] # Compare scores

      # Print scores for decision tree
      print("Decisiontree Scores")
      print(grid_search_cv.best_params_)
      print(grid_search_cv.best_score_)
      print("-------------------------")

      # Print scores for logistic regression
      print("LogisticRegression Scores")
      print(logistic_grid.best_params_)
      print(logistic_grid.best_score_)
      print("-------------------------")
```

```
Decisiontree Scores
{'criterion': 'entropy'}
0.8243274205469326
-------------------------
LogisticRegression Scores
{'C': 0.01, 'penalty': 'l2'}
0.8127753141167775
-------------------------
```

- Decision Tree > Logistic Regression
- This model gave us the best result.

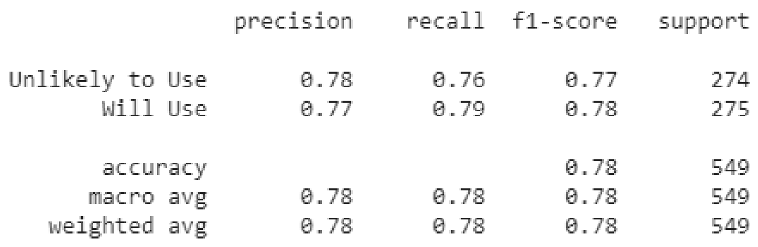
Model Results comparison

- Decision Tree Model Results



```
                precision    recall  f1-score   support

Unlikely to Use      0.80      0.78      0.79       274
       Will Use      0.79      0.80      0.79       275

       accuracy                          0.79       549
      macro avg      0.79      0.79      0.79       549
   weighted avg      0.79      0.79      0.79       549
```

- Logistic Regression Model Results



```
                precision    recall  f1-score   support

Unlikely to Use      0.78      0.76      0.77       274
       Will Use      0.77      0.79      0.78       275

       accuracy                          0.78       549
      macro avg      0.78      0.78      0.78       549
   weighted avg      0.78      0.78      0.78       549
```

- Decision Tree > Logistic Regression
- This model gave us the best result.

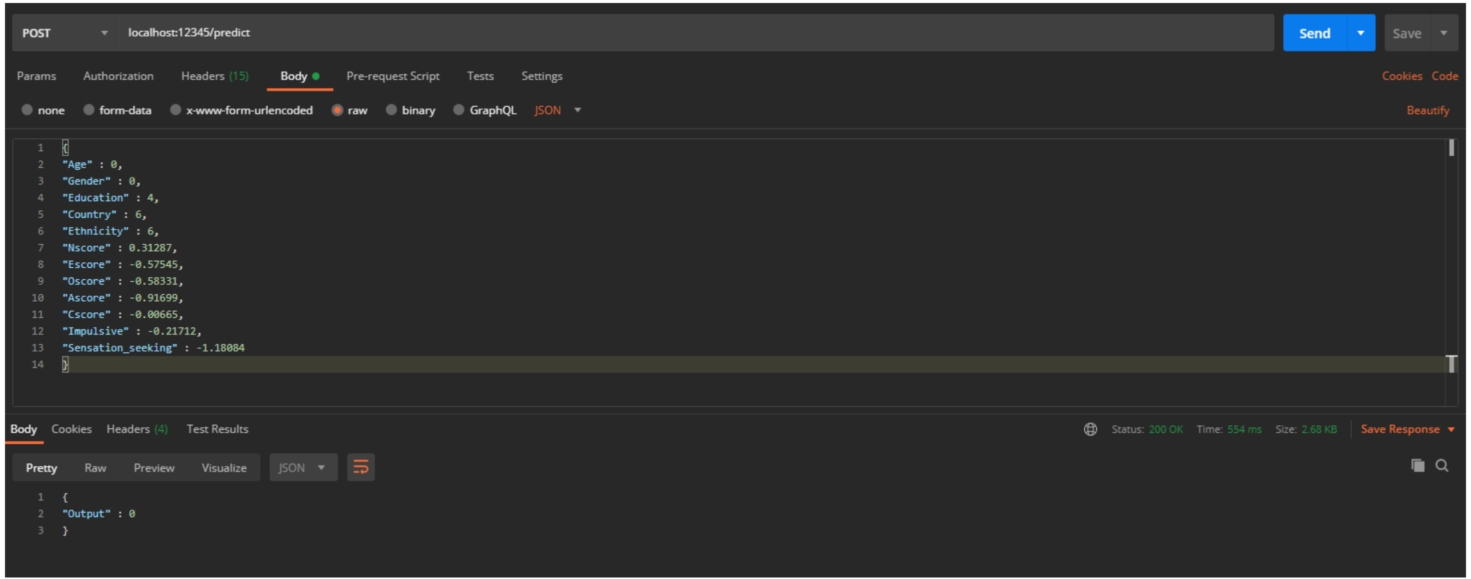
API

- **API** is the acronym for Application Programming Interface, which is a software intermediary that allows two applications to talk to each other.
- We defined in our API a method that takes in parameter individual's attributes and from it, it predicts whether the individual will be a cannabis user or not.
- **Flask** is a web framework. This means flask provides you with tools, libraries and technologies that allow you to build a web application.

API Flask

- Passing a **JSON** via a **POST request** in the following format :

```
{
"Age" : Number,
"Gender" : Number,
"Education" : Number,
"Country" : Number,
"Ethnicity" : Number,
"Nscore" : Number,
"Escore" : Number,
"Oscore" : Number,
"Ascore" : Number,
"Cscore" : Number,
"Impulsive" : Number,
"SS" : Number
}
```



POST localhost:12345/predict

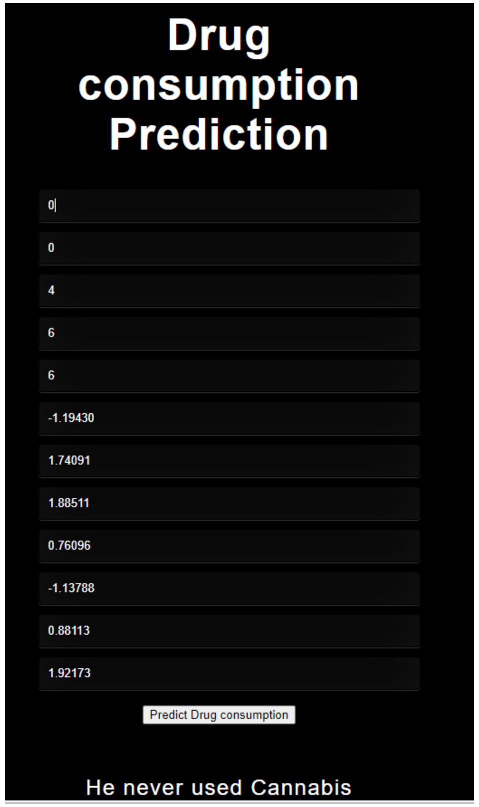
```
{
"Age" : 0,
"Gender" : 0,
"Education" : 4,
"Country" : 6,
"Ethnicity" : 6,
"Nscore" : 0.31287,
"Escore" : -0.57545,
"Oscore" : -0.58331,
"Ascore" : -0.91699,
"Cscore" : -0.00665,
"Impulsive" : -0.21712,
"Sensation_seeking" : -1.18084
}
```

Status: 200 OK Time: 554 ms Size: 2.68 KB

```
{
"Output" : 0
}
```

API Flask

- We also tried to do a web interface, but unfortunaly it doesn’t work.



Drug consumption Prediction

0

0

4

6

6

-1.19430

1.74091

1.88511

0.76096

-1.13788

0.88113

1.92173

He never used Cannabis