Cannabis consumption prediction from demographics and personality

HarvardX PH125.9x Data Science Capstone

Charles Mégnin

10/2/2019

Executive Summary

- Goal of project

The goal of this project is to assess whether an individual's consumption of cannabis can be predicted from a combination of demographic and personality data.

- Dataset description

The original dataset is found on the UCI machine learning repository. It is based the research paper by E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," arXiv, 2015. The data was collected from 1885 English-speaking participants over 18 years of age between March 2011 and March 2012.

In the original dataset, drug use is separated between 'Never used', 'Used over a decade ago', 'Used over a decade ago', 'Used in last decade', 'Used in last year', 'Used in last month', 'Used in last week' and 'Used in last day'. For the purpose of this study, we separate the data in two groups: 'Never Used' (the original predictor) and 'Used' (the combination of the others).

The original dataset includes questions related to the use of alcohol, amphetamines, amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, magic mushrooms, nicotine and volatile substance abuse (VSA)) and one fictitious drug (Semeron) which was introduced to identify over-claimers. In the present study, we restrict our scope to examining cannabis consumption.

The data is divided into five demographic predictors : Age, Gender, Level of education, Ethnicity and Country of origin.

In addition, the data includes the results from seven tests administered to assess personality, specifically:

1. Neuroticism (a long-term tendency to experience negative emotions such as nervousness, tension, anxiety and depression);
2. Extraversion (manifested in outgoing, warm, active, assertive, talkative, cheerful, and in search of stimulation characteristics);
3. Openness to experience (a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests);
4. Agreeableness (a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness);
5. Conscientiousness (a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient);
6. Impulsiveness;
7. Sensation-seeking.

The working dataset in this study therefore consists of one Class (Cannabis consumption labeled 'Used') and twelve predictors (five demographic and seven personality-related).

- Key steps

We extract a training subset (80% of data) from the dataset for the purpose of training our model, and use the remaining 20% of the data as a test set for the purpose of evaluating the goodness of fit of our model. it being a classification problem, we use the accuracy as the metric for fit with the test set.

This report consists of two main sections:

- In the first part, after performing minor data engineering, we explore and analyze the dataset.
- In the second part, we move on to the modeling phase.

Analysis

A: Data engineering and checks

All predictors have already been normalized and centered by the authors of the original paper.

We construct the ‘Used’ class to separate ‘Never used’ which we label “0”" from the others which we label “1”.

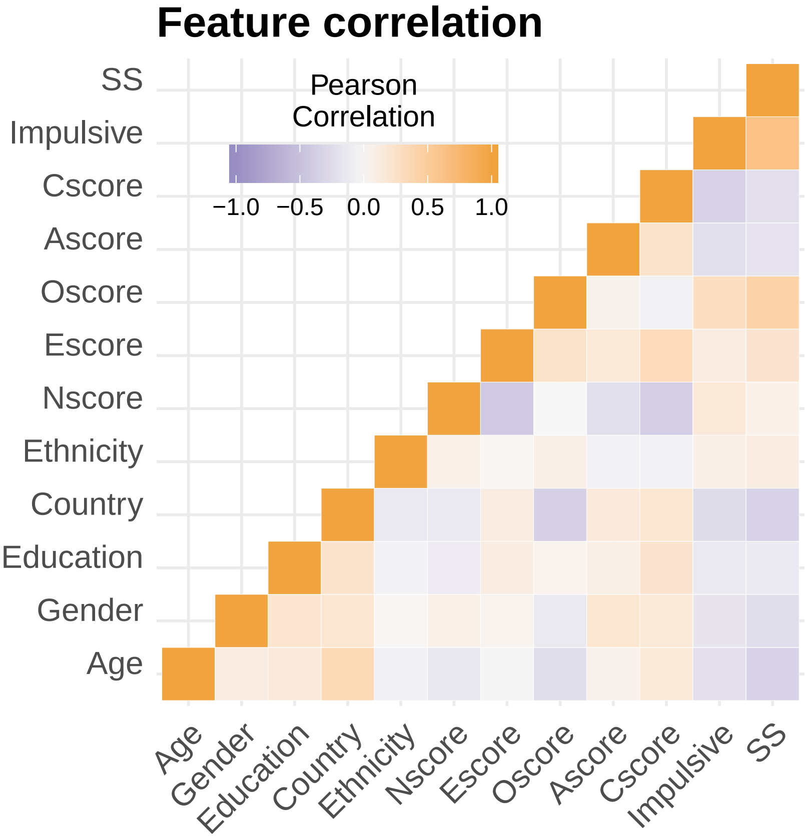
There are 0 NAs in the dataset.

We then partition the data between training (80%) and test sets (20%) preserving the distribution of the Cannabis class.

B: Data exploration

Feature correlation analysis

We examine whether any redundancies are present among the predictors:

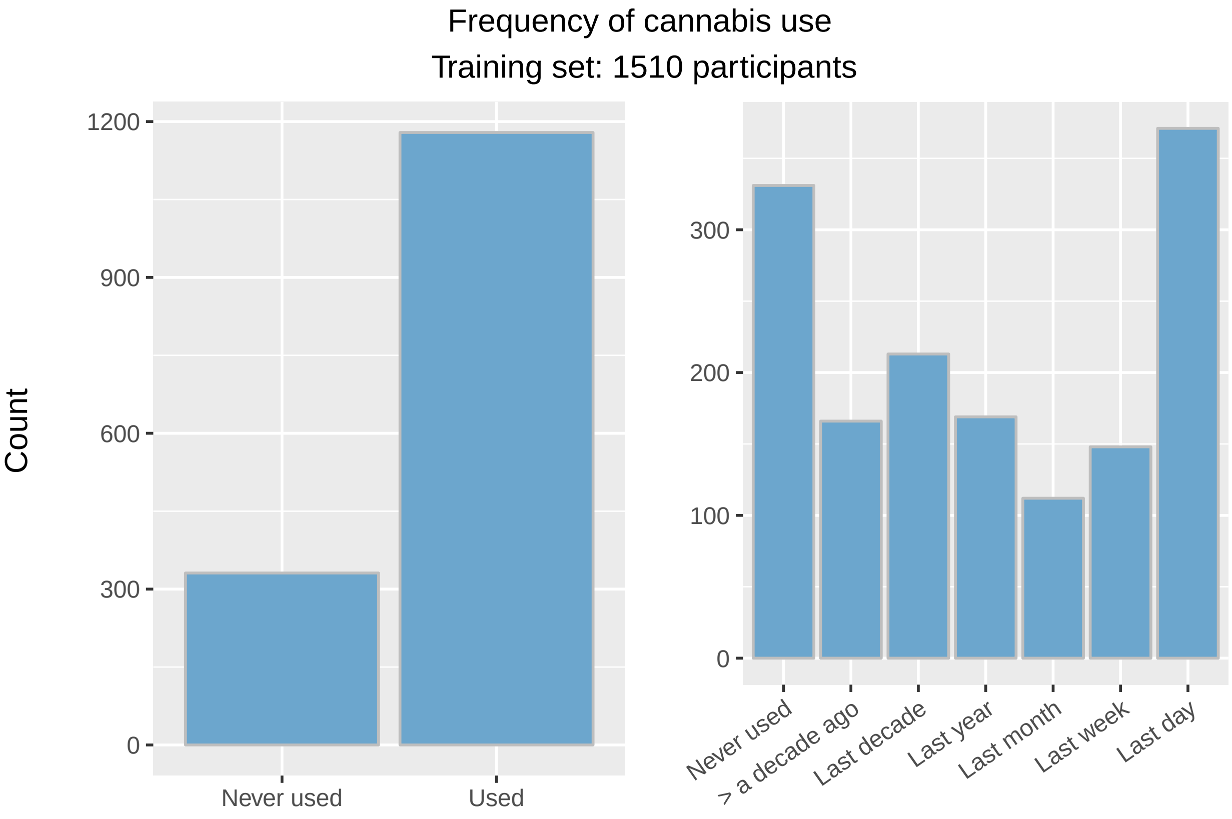


Most features are weakly inter-correlated, the strongest correlation is between Impulsivity and Sensation-seeking (0.622848 correlation, p-value = 0). It is safe to use all features for modeling purposes.

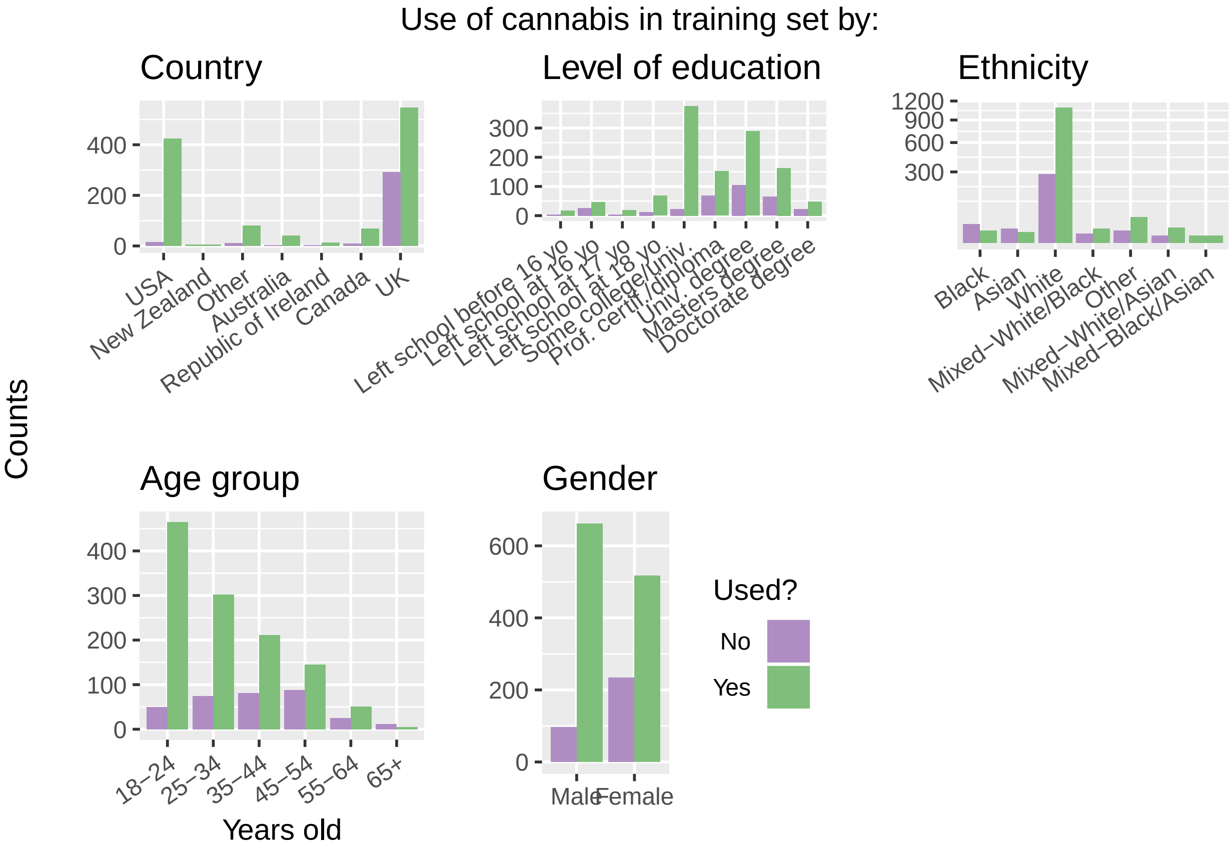
Data exploration

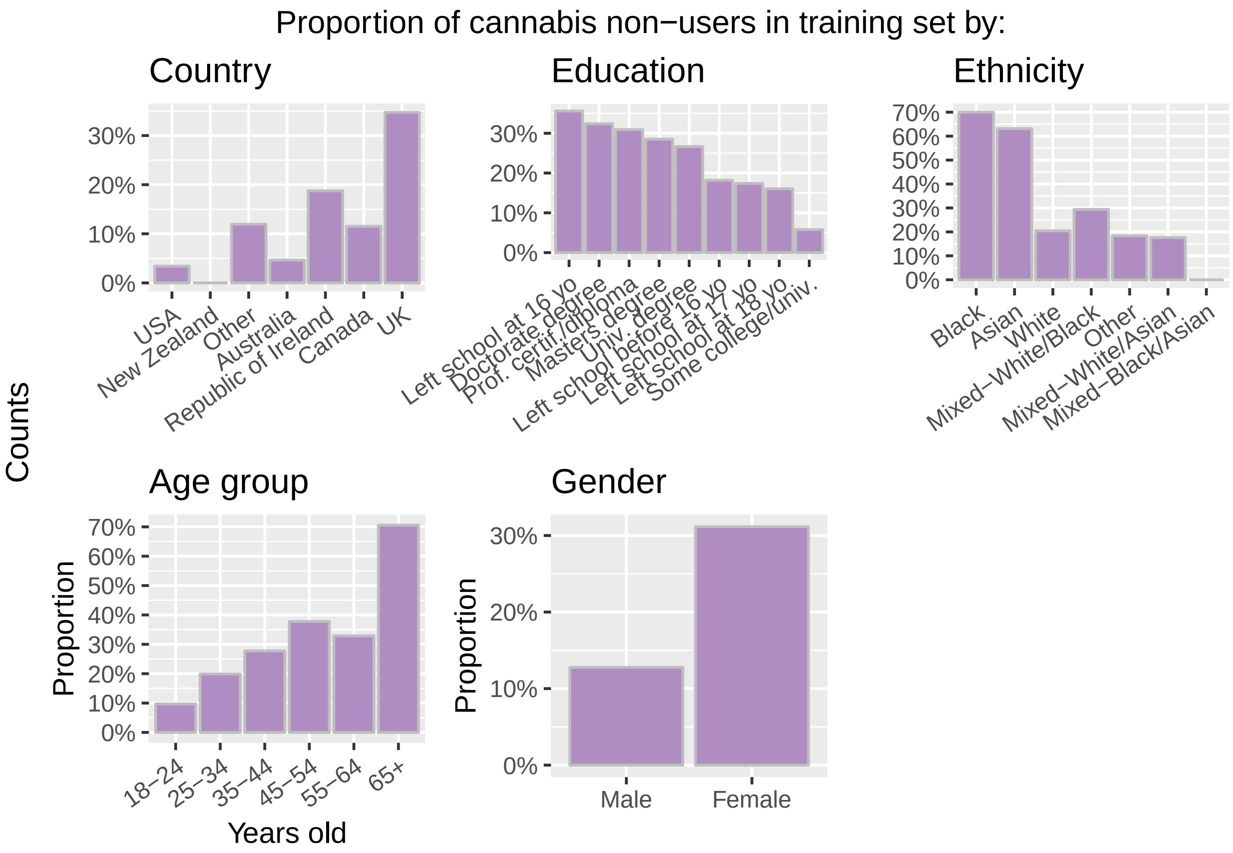
Overall cannabis consumption

The training set of 1510 participants (left plot) is binned into 1179 users and 331 non-users. The original data is shown on the right.

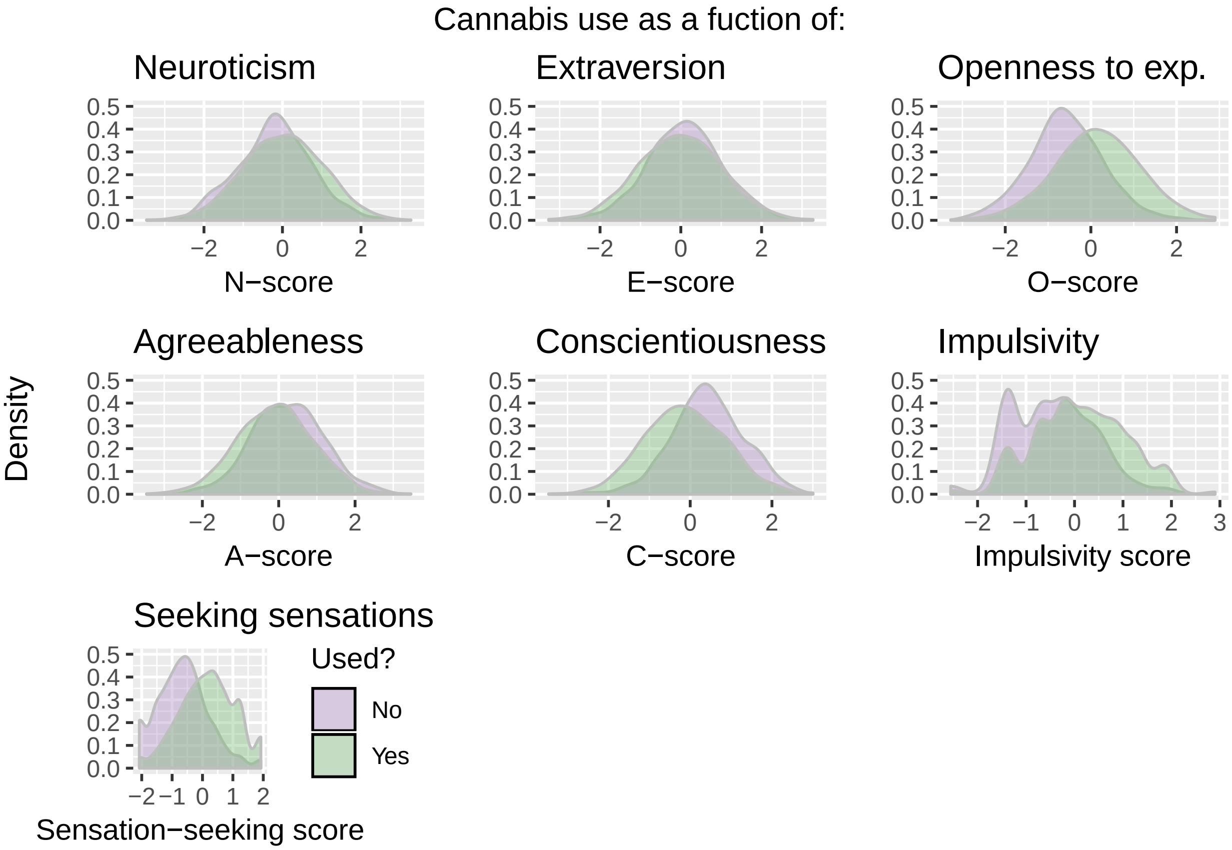


Demographic analysis




Proportion of cannabis non−users in training set by:
Country
30%
20%
10%
0%
USA
New Zealand
Other
Australia
Republic of Ireland
Canada
UK
Education
30%
20%
10%
0%
Left school at 16 yo
Doctorate degree
Prof. certif./diploma
Masters degree
Univ. degree
Left school before 16 yo
Left school at 17 yo
Left school at 18 yo.
Some college/univ.
Ethnicity
70%
60%
50%
40%
30%
20%
10%
0%
Black
Asian
White
Mixed−White/Black
Other
Mixed−White/Asian
Mixed−Black/Asian
Counts
Age group
Proportion
70%
60%
50%
40%
30%
20%
10%
0%
18−24
25−34
35−44
45−54
55−64
65+
Years old
Gender
Proportion
30%
20%
10%
0%
Male
Female

Personality analysis



The density plots show some measure of difference between users and non-users particularly as it relates to openness to experience, agreeableness, conscientiopusness, impuslivity and seeeking sensations. However, Neuroticism and extraversion do not seem to play much of a role. We will examine in the modeling section whether that is indee the case.

C: Modeling

Results

Conclusion