Prediction of Cannabis Consumption from Demographics and Personality

HarvardX PH125.9x Data Science Capstone

Charles Mégnin

10/2/2019

Contents

Executive Summary

- Introduction

Drug use is a behavior that constitutes an important factor linked to poor health, including early mortality, and which presents significant adverse consequences for the social fabric, notably with respect to criminality and family cohesion. Early detection of an individual's predisposition to drug consumption offers the opportunity to healthcare professionals to short-circuit the onset of addiction.

The present study is based on a dataset that includes demographic and psychological information related to the consumption of 18 legal and illegal drugs by 1885 participants. For the purpose of this study, we choose to focus the data analysis and modeling to the use of cannabis.

- 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.

To do so, we build and assess the effectiveness of a number (***HOW MANY?***) of machine learning classifiers and confront the results obtained to the insights provided by data exploration.

- 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 consists of two groups of pre-normalized and centered predictors:

1. Five demographic predictors : Age, Gender, Level of education, Ethnicity, and Country of origin.
2. The results from seven scored tests administered to assess personality, specifically:

- Neuroticism (a long-term tendency to experience negative emotions such as nervousness, tension, anxiety and depression);
- Extraversion (manifested in outgoing, warm, active, assertive, talkative, cheerful, and in search of stimulation characteristics);
- Openness to experience (a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests);
- Agreeableness (a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness);
- Conscientiousness (a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient);
- Impulsiveness;
- 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 each classifier. This being a classification problem, we use the accuracy as the metric to assess the goodness of fit.

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:
- After applying a Recursive feature elimination algotirhm to remove predictors that do not contribute significantly to the outcome, we build models based on the following methods:
- Generalized linear model
- Decision tree
- Random forest

- Stochastic gradient boosting

We compare the modeling approaches, both in terms of accuracy and coherence with the data analysis.

Analysis

A: Data engineering

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

We construct the 'Used' class to separate 'Never used' participants which we label "0"" from the others which we label "1".

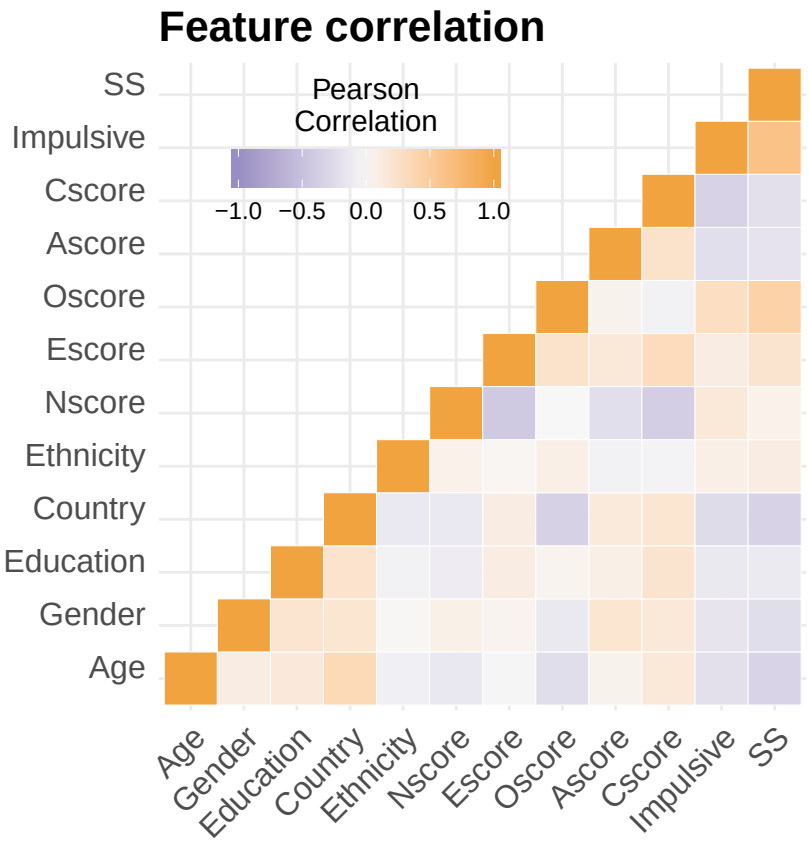
There are 0 NAs in the dataset.

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

B: Data exploration

Feature correlation

We examine whether any redundancies are present among the 12 predictors:

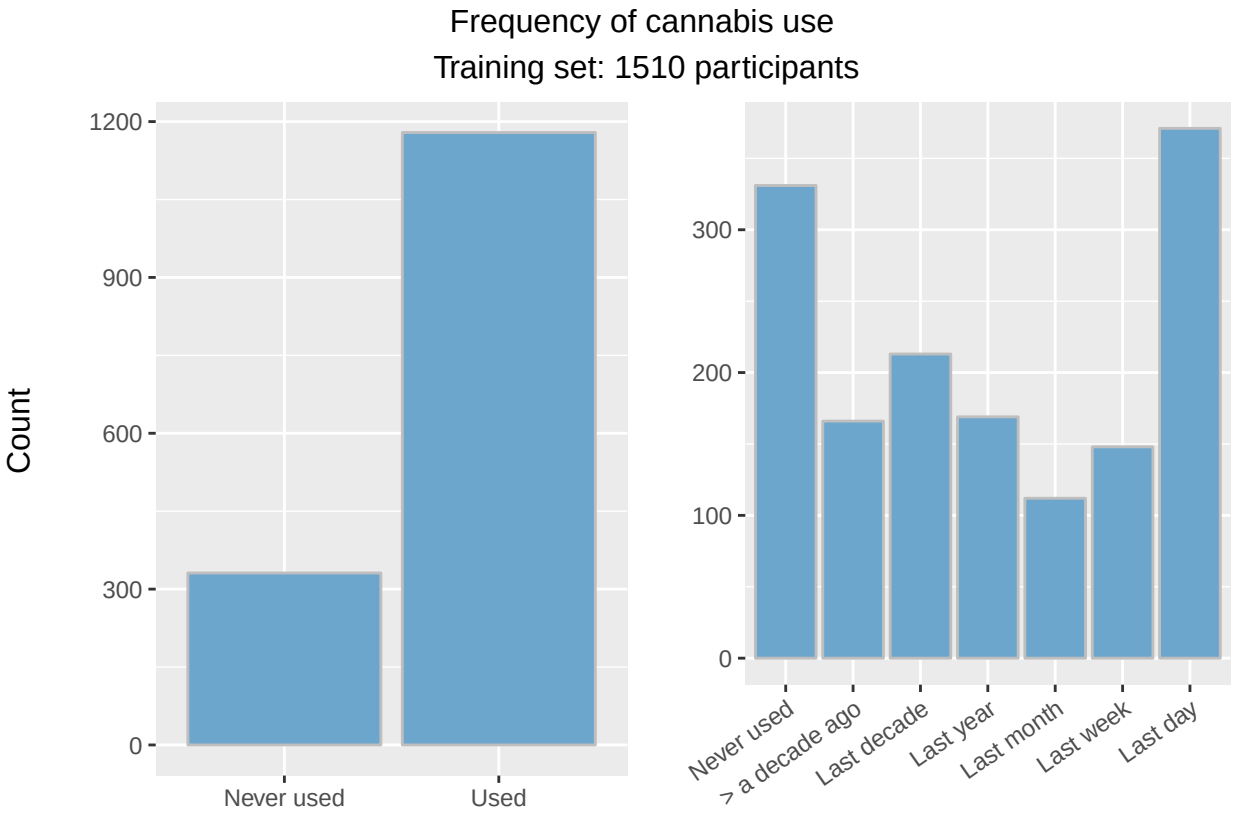


Most features are weakly inter-correlated, the strongest correlation is between Impulsivity and Sensation-seeking (0.622848 correlation, p-value = 0): there are no redundancies among features for modeling purposes.

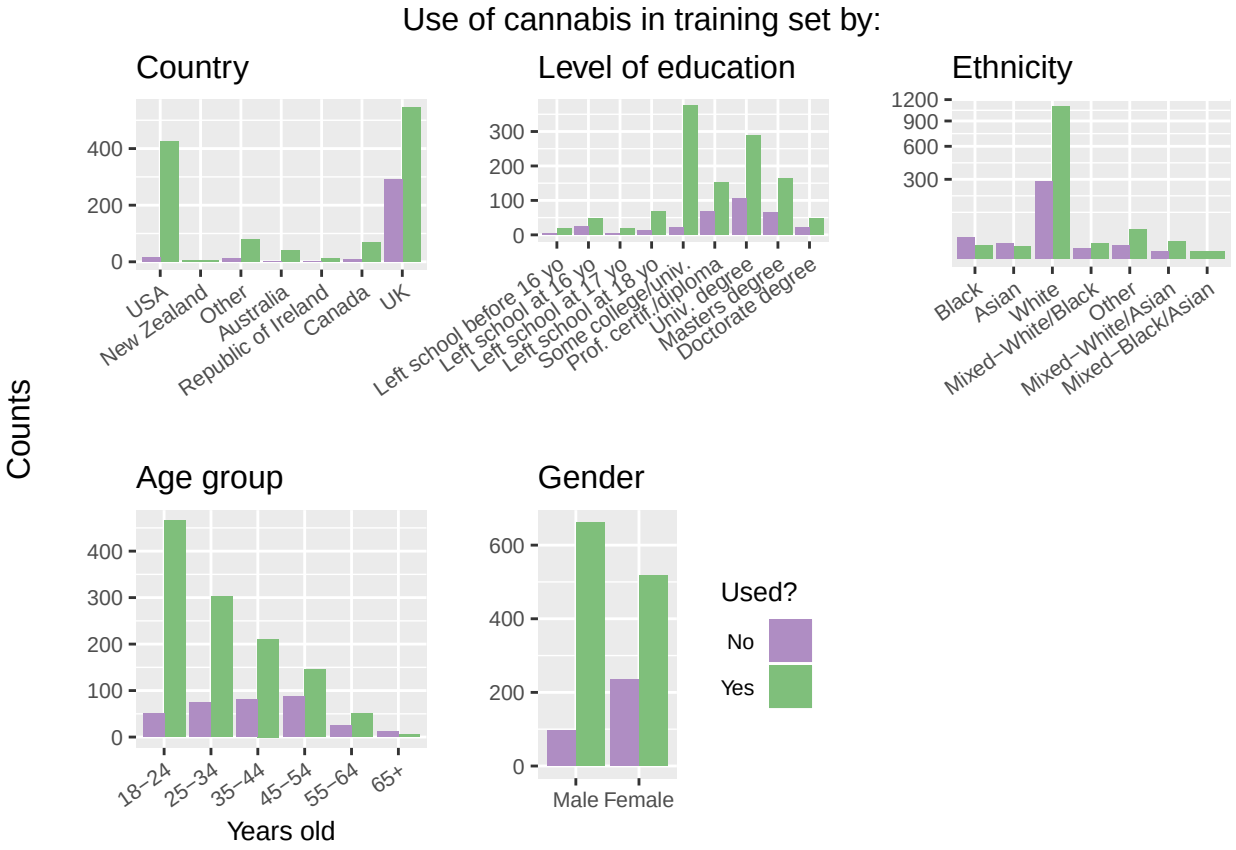
Data exploration

Overall cannabis consumption

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

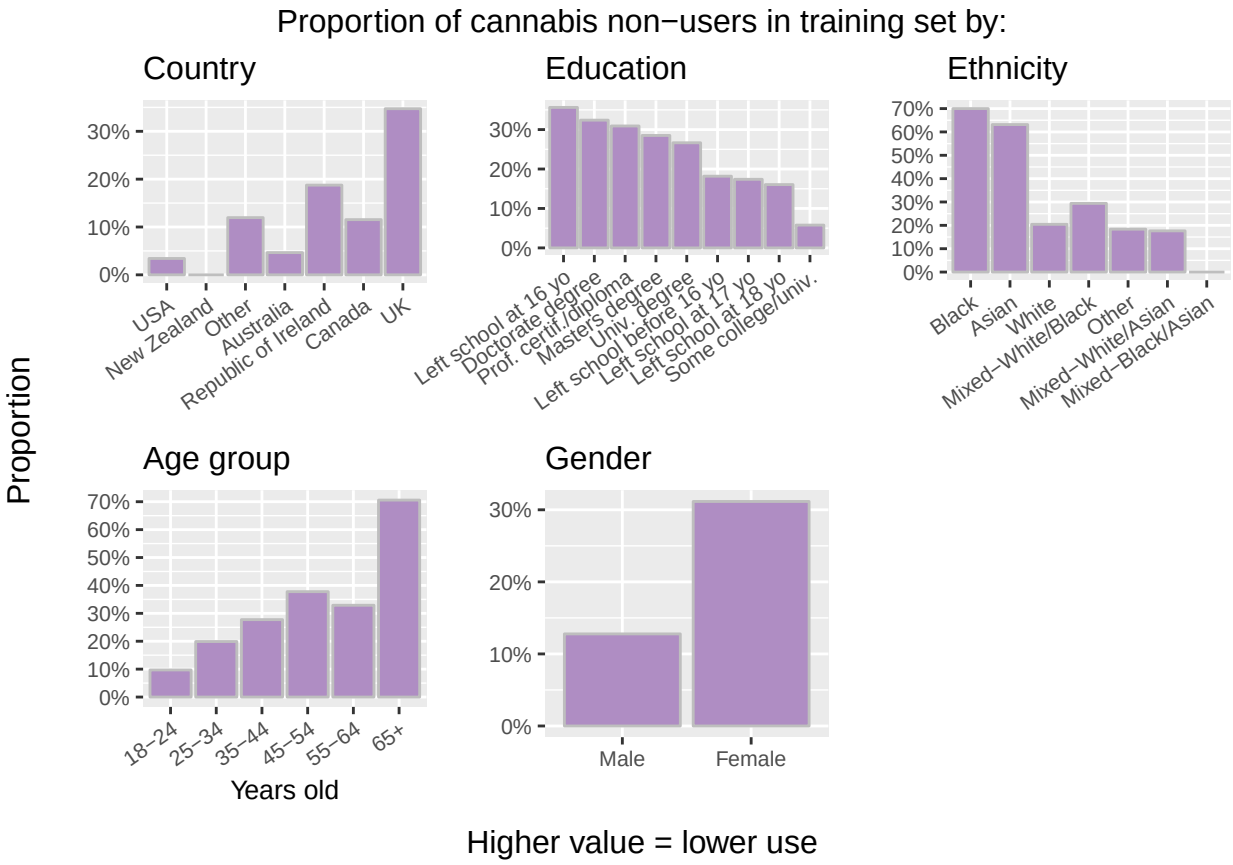


Demographic analysis (before re-binning)



The dataset is dominated by young and educated white American and British participants of both sexes. The small sizes of many sub-groups (New Zealanders (4), Irish (16), Canadians (78), those that left school as teenagers, Mixed black-asians (3), and people over 65 years of age(17)) to name just a few adds little valuable insight and will likely only serve to introduce variance in the analysis.

Given the preponderance of cannabis users in the dataset, examining the proportion of non-users in the population is more insightful than that of users:



The figure shows:

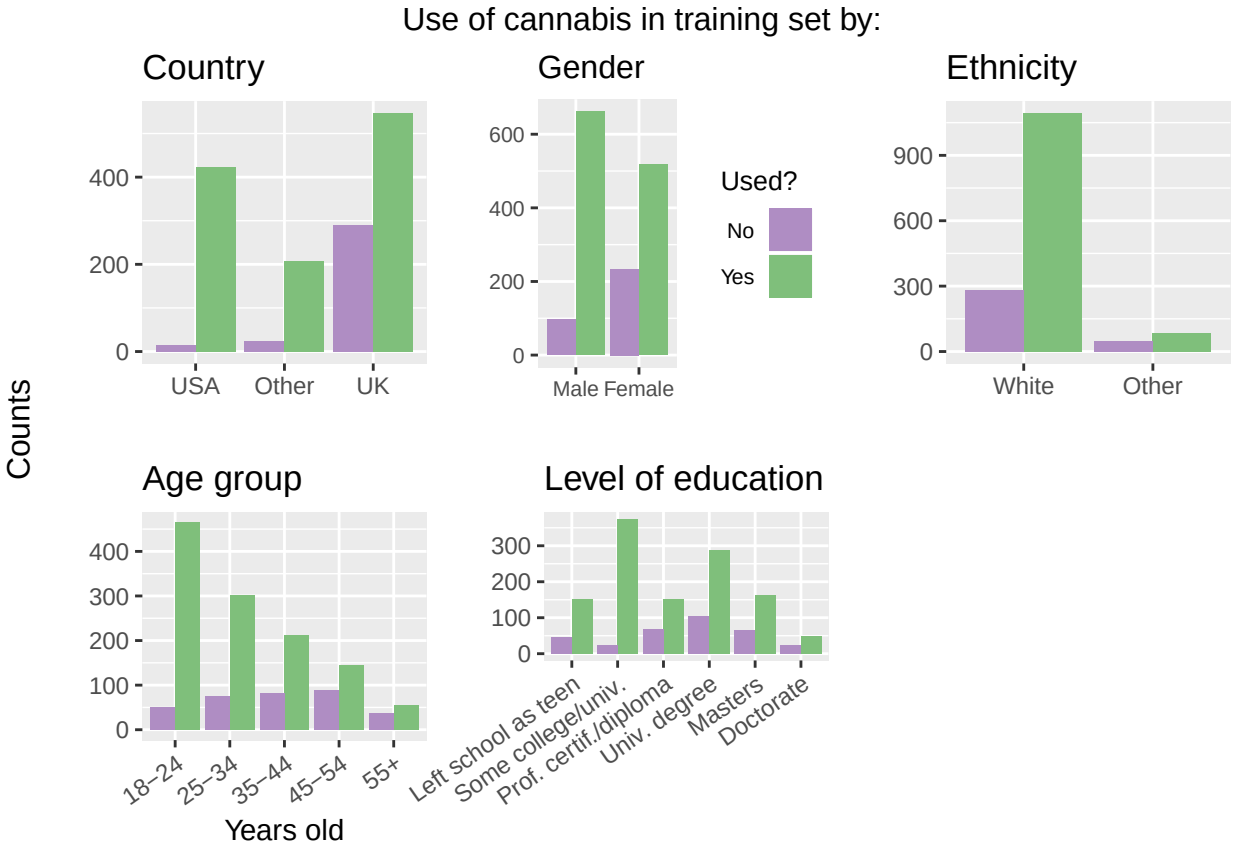
- British participants have the lowest propensity at using cannabis (65.274463%) and New Zealanders the highest (100%). The latter observation may be little more than the manifestation of the high variance resulting from the small sample size (4 participants from New Zealand over a total of 1510).
- While the most educated participants are less likely to consume, those that left school at 16 years old consume least of all. As in the case of New Zealanders, the latter observation may be an artifact of the small sample size (73 participants left school at 16 years old).
- The data shows that, of all ethnic groups, those that identified as blacks and Asians consume the least cannabis (30% and 36.8421053% respectively) compared to, say, whites (79.5636364%). Nevertheless the dataset is ethnically dominated by whites (1375 samples), and here again, results for blacks (30 samples), as well as Asians (19 samples) and other minorities under-represented in the dataset may not be meaningful.
- Age seems to be highly correlated with cannabis consumption, with use decreasing almost steadily with the age group. Given that we only distinguish between people having never used from the others following a 'virginity from cannabis' perspective, this points to a generational phenomenon. Perhaps unsurprisingly, we note a large discrepancy between the generations that precede and those that follow the 1950s.
- Females are 2.4380688 times less likely to consume cannabis than males.

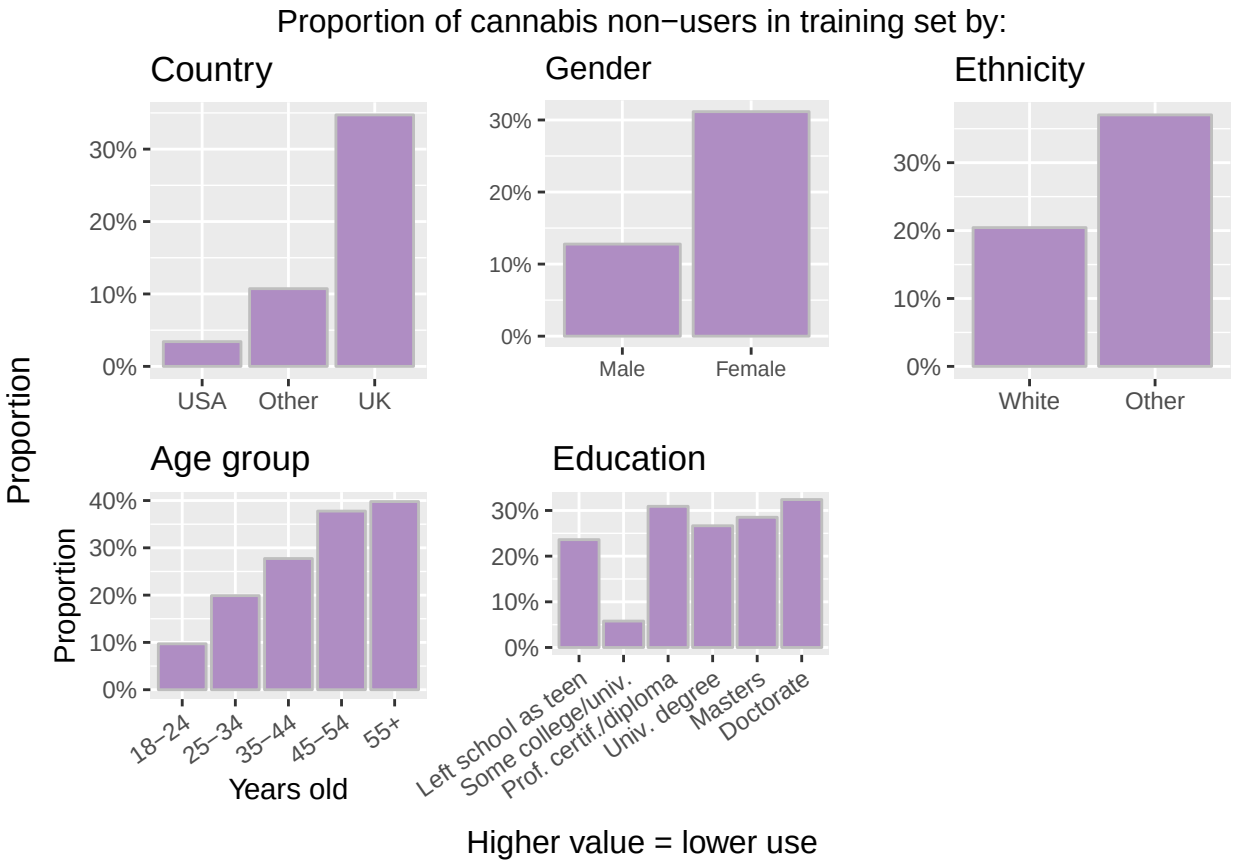
At the risk of erasing behavioral differences among groups, the distribution of the dataset forces a more realistic binning of the demographic information:

- 3 groups for Country: "USA", "UK", "Others".
- 6 groups for Education: "Left school as a teen", "Some college", "Professional certificate", "University degree", "Masters degree", "Doctorate degree".

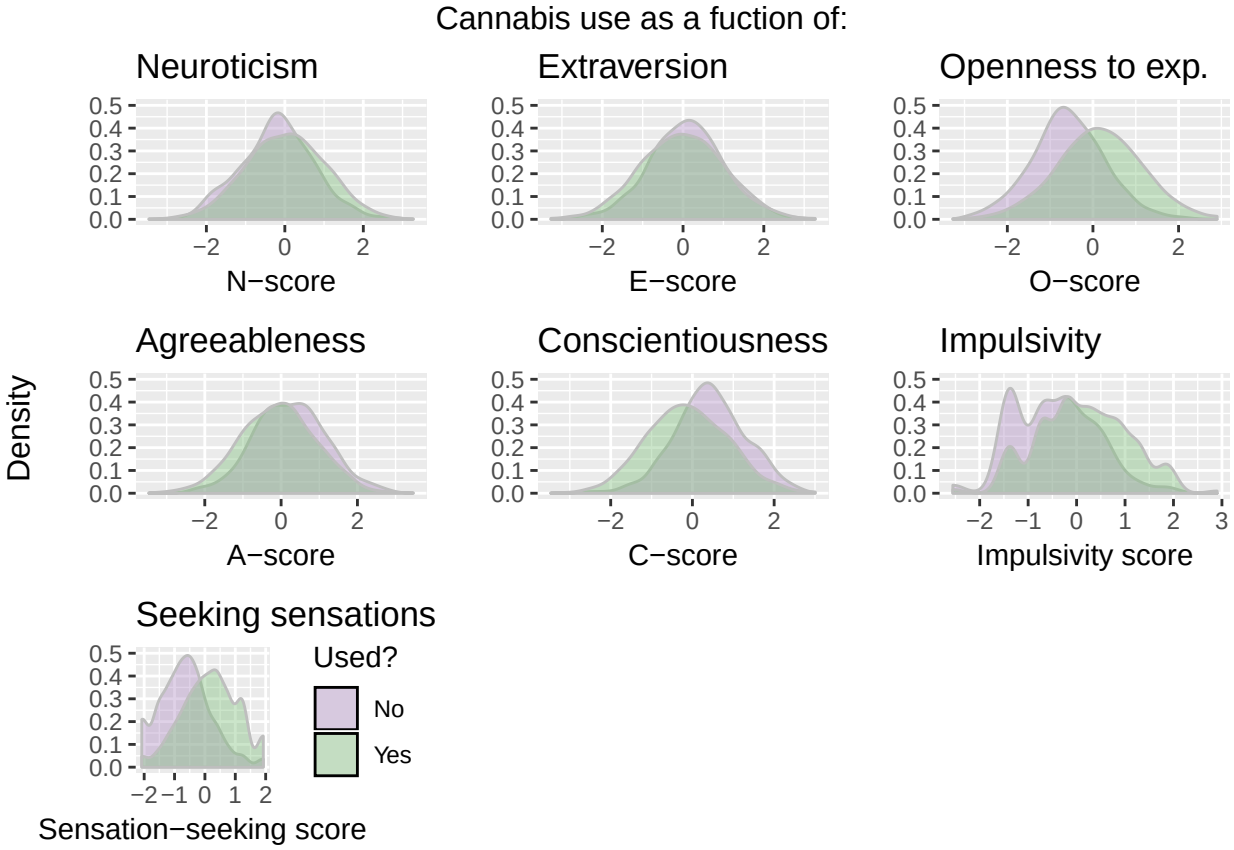
- 2 ethnic groups: “Whites”, “Non-whites”
- 5 age groups: “18-24”, “25-34”, “35-44”, “45-54”, “55+”

Demographic analysis (after re-binning)




Proportion of cannabis non−users in training set by:
Country
Gender
Ethnicity
Age group
Education
Proportion
Years old
USA
Other
UK
Male
Female
White
Other
18–24
25–34
35–44
45–54
55+
Left school as teen
Some college/univ.
Prof. certif./diploma
Univ. degree
Masters
Doctorate
Higher value = lower use

Personality analysis



The density plots show some measure of difference between users and non-users particularly as it relates to openness to experience, agreeableness, conscientiousness, impulsivity, and seeking sensations. However, it is difficult to distinguish users from non-users when it comes to neuroticism and extraversion. We will examine in the modeling section below whether that is indeed the case.

C: Modeling

We seek a model which improves on the ratio of users to the population (0.7807947). This constitutes the baseline above which predictive modeling becomes interesting.

Recursive Feature elimination

For RFE as well as subsequent modeling, we use the k-fold cross validation method which involves splitting the dataset into k subsets. The algorithm holds aside one of the subsets while the model is trained on the others. This process is repeated a predetermined number of times and the overall accuracy estimate is provided.

Results

Conclusion

Noise introduced by the segmentation of data into very small groups.