Prediction of Drug Consumption Based on Five Factor Personality Features

Eric Gabriel

January 5th, 2021

Contents

Abstract

In this report, we performed binary classification of users and non-users based on the Drug Consumption (Quantified) data set published by (Fehrman et al. 2015). This classification can be used to assess the risk of an individual being a user of a specific drug based on personality features. After an exploratory data analysis (EDA), we decided on three drugs and four different classes of models for evaluation.

The best performance was achieved using regularized discriminant analysis (RDA) models: We report balanced accuracies of 66.5%, 79.2% and 73.2% for the drugs Benzodiazepines, Legal Highs and Ecstasy, respectively. The RDA model outperforms the results reported by (Mahyoub et al. 2019) and performs slightly worse than the highly optimized decision trees of (Fehrman et al. 2015).

A potential improvement is demonstrated by experimental results when using additional input features exploiting drug correlation. Further improvements are expected by using an optimal input feature sub-set for each drug as well as a more sophisticated tune grid for the used model types.

Introduction

This report is part of the second project submission of the course **HarvardX PH125.9x "Data Science: Capstone"**.

Over a very large period of time, the medical advancement has led to a large variety of available drugs to cure or ease an even larger variety of diseases. On the downside, drug abuse is – in most cases – as old as the respective drug itself. Following (Kleiman, Caulkins, and Hawken 2011), a drug is a chemical substance that impacts biological functionality apart from providing hydration or nutrition. The abuse of drugs, i.e., any use that is not medically indicated, puts the health of the respective individual at risk. Regarding legal drugs as alcohol and tobacco (nicotine), a negative impact on life expectancy can not be neglected (Beaglehole et al. 2011). For many other drugs, the effect can be even worse.

The ability to assess the risk of an individual abusing drugs and/or becoming a regular user based on personality data can highly support finding the right prevention strategy and to address the right target groups.

The goal of this report is to generate and compare models that are able to predict a person being a user or non-user of a specific drug based on personality data.

This report is structured as follows: We first describe the data set and perform an exploratory data analysis (EDA). Afterwards, three drugs are chosen for which we will train different classes of models. We will use insights from the EDA to choose appropriate models. The methodology, the experimental setup and the evaluation metrics are explained afterwards. In the following section, we present the results of the trained classifiers. The comparison between different models, different drugs and results from the literature is then discussed. After we present ideas for further improvements with one detailed example, the report closes with a summary.

Problem Statement

The problem to be solved is the prediction of a person being a user or non-user of a specific drug on a given data set: Drug Consumption (Quantified). These predictions will be generated using different models that operate on a certain set of input features. We will use different types of models, different drugs and also other results from the literature for performance comparison.

More specifically, the problem at hand is a binary classification task per examined drug type with supervised learning.

Data Set: Drug Consumption (Quantified)

To classify users and non-users, the data set **Drug Consumption (Quantified)** (Fehrman et al. 2015) is used. It can be downloaded from the UCI Machine Learning Repository ("Drug Consumption (Quantified) at UCI Machine Learning Repository" 2016).

The data set was collected between March 2011 and March 2012 using an anonymous online survey tool. After an input validation, $1,885$ observations remain.

The survey is split into questions regarding personal background, questions on personality information as well as questions about the use and recency of 18 legal and illegal drugs (+ one fictitious drug to identify over-claimers):

Personal Information

- **Age**: 18 - 24, 25 - 34, 35 - 44, 45 - 54, 55 - 64, 65+
- **Gender**: Male, Female

- **Education**: Left school before 16y, Left school at 16y, Left school at 17y, Left school at 18y, Some college or university (no certificate or degree), Professional certificate / Diploma, University degree, Master's Degree, Doctorate degree
- **Country**
- **Ethnicity**

Personality Information

- 60 questions of **Revised NEO Five Factor Inventory** (NEO-FFI-R) (Costa and MacCrae 1992), 12 per factor, 0 - 5 points each (strongly disagree - strongly agree):

1. Neuroticism (**N**): Long-term tendency for negative emotions
2. Extraversion (**E**): Outgoing, talkative, cheerful etc.
3. Openness to Experience (**O**): Creative, imaginative, unusual ideas, art
4. Agreeableness (**A**): Trust, modesty, kindness etc.
5. Conscientiousness (**C**): Organized, dependable, reliable, efficient etc.

- 30 questions of Barrat Impulsiveness Scale (BIS-11) (Stanford et al. 2009) to measure motor and attentional impulsiveness.

- 30 questions of Impulsiveness Sensation-Seeking (ImpSS) (Zuckerman 1994): 19 true/false-statements to measure impulsiveness and 11 items measuring sensation-seeking.

Drug Use

The data set contains one of the following statements per individual for each of the 18 legal and illegal drugs (plus one fictitious drug) about the recency of use (if applicable):

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

The 19 drugs are the following:

- Alcohol
- Amphetamines
- Amyl nitrite
- Benzodiazepine
- Cannabis
- Chocolate
- Cocaine
- Caffeine
- Crack
- Ecstasy
- Heroin
- Ketamine
- Legal highs
- LSD
- Methadone
- Magic Mushrooms
- Nicotine
- Volatile substance abuse (VSA)
- Semeron (fictitious drug to identify over-claimers)

Quantification

All categorical features have been quantified by (Fehrman et al. 2015), i.e., transformed into numerical features.

The five scores of NEO-FFI-R have been converted into sample-based T-Scores with mean 50 and standard deviation 10 using:

$$\text{T-Score}_{\text{sample}} = 10\left[\frac{\text{Raw score} - \text{Sample mean}}{\text{Sample standard deviation}}\right] + 50 \tag{1}$$

The sample mean was chosen because t-tests have shown that the sample mean deviates significantly from the population mean. The use of sample-based T-Scores supports the analysis of classifying individuals into users and non-users in this data set. The reader is referred to (Fehrman et al. 2015) for further details.

Ordinal features were quantified using polychloric correlation (Lee, Poon, and Bentler 1995) (Martinson and Hamdan 1971). Here, the average probability is used for each interval.

Nominal features such as gender and ethnicity were quantified using a non-linear categorical principal component analysis (CatPCA) (Linting and van der Kooij 2012).

Afterwards, all quantified features have been normalized such that they have a mean of 0 and a standard deviation of 1. Please see (Fehrman et al. 2015) for further details.

Data Preparation and Pre-Processing

The downloaded data set file is in CSV format and contains no header lines. Thus, we set the respective column names before displaying the data.

Additionally, we convert the 19 columns regarding the use of drugs into a numeric representation:

- Never Used (0)
- Used over a Decade Ago (1)
- Used in Last Decade (2)
- Used in Last Year (3)
- Used in Last Month (4)
- Used in Last Week (5)
- Used in Last Day (6)

In order to perform binary classification per drug, (Fehrman et al. 2015) suggest to call an individual a user if the drug was used within the last decade or more recently. This corresponds to a value $>$ 1. Otherwise, the individual is a non-user. Following this approach, we add one column for each drug ("<DRUG_NAME>_User“) with two factors each (”yes" and “no”) indicating if the individual is a user or a non-user, respectively.

```
# Print the summary of the data set
summary(data_clean)
```

```
##        ID              Age                Gender             Education
##  Min.   :   1.0   Min.   :-0.95197   Min.   :-0.4824600   Min.   :-2.435910
##  1st Qu.: 474.0   1st Qu.:-0.95197   1st Qu.:-0.4824600   1st Qu.:-0.611130
##  Median : 946.0   Median :-0.07854   Median :-0.4824600   Median :-0.059210
##  Mean   : 945.3   Mean   : 0.03461   Mean   :-0.0002559   Mean   :-0.003806
##  3rd Qu.:1417.0   3rd Qu.: 0.49788   3rd Qu.: 0.4824600   3rd Qu.: 0.454680
##  Max.   :1888.0   Max.   : 2.59171   Max.   : 0.4824600   Max.   : 1.984370
##     Country          Ethnicity           NScore              EScore
##  Min.   :-0.5701   Min.   :-1.1070   Min.   :-3.464360   Min.   :-3.273930
##  1st Qu.:-0.5701   1st Qu.:-0.3169   1st Qu.:-0.678250   1st Qu.:-0.695090
##  Median : 0.9608   Median :-0.3169   Median : 0.042570   Median : 0.003320
```

```
##  Mean   : 0.3555   Mean   :-0.3096   Mean   : 0.000047   Mean   :-0.000163
##  3rd Qu.: 0.9608   3rd Qu.:-0.3169   3rd Qu.: 0.629670   3rd Qu.: 0.637790
##  Max.   : 0.9608   Max.   : 1.9072   Max.   : 3.273930   Max.   : 3.273930
##      OScore              AScore              CScore
##  Min.   :-3.273930   Min.   :-3.464360   Min.   :-3.464360
##  1st Qu.:-0.717270   1st Qu.:-0.606330   1st Qu.:-0.652530
##  Median :-0.019280   Median :-0.017290   Median :-0.006650
##  Mean   :-0.000534   Mean   :-0.000245   Mean   :-0.000386
##  3rd Qu.: 0.723300   3rd Qu.: 0.760960   3rd Qu.: 0.584890
##  Max.   : 2.901610   Max.   : 3.464360   Max.   : 3.464360
##  Impulsiveness       SensationSeeking       Alcohol       Amphetamine
##  Min.   :-2.555240   Min.   :-2.078480   Min.   :0.000   Min.   :0.000
##  1st Qu.:-0.711260   1st Qu.:-0.525930   1st Qu.:4.000   1st Qu.:0.000
##  Median :-0.217120   Median : 0.079870   Median :5.000   Median :0.000
##  Mean   : 0.007216   Mean   :-0.003292   Mean   :4.635   Mean   :1.341
##  3rd Qu.: 0.529750   3rd Qu.: 0.765400   3rd Qu.:6.000   3rd Qu.:2.000
##  Max.   : 2.901610   Max.   : 1.921730   Max.   :6.000   Max.   :6.000
##   AmylNitrite          Benzos          Caffeine        Cannabis
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:5.000   1st Qu.:1.000
##  Median :0.0000   Median :0.000   Median :6.000   Median :3.000
##  Mean   :0.6069   Mean   :1.465   Mean   :5.484   Mean   :2.989
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:6.000   3rd Qu.:5.000
##  Max.   :6.0000   Max.   :6.000   Max.   :6.000   Max.   :6.000
##    Chocolate        Cocaine          Crack           Ecstasy
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:5.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000
##  Median :5.000   Median :0.000   Median :0.0000   Median :0.000
##  Mean   :5.107   Mean   :1.161   Mean   :0.2976   Mean   :1.314
##  3rd Qu.:6.000   3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:3.000
##  Max.   :6.000   Max.   :6.000   Max.   :6.0000   Max.   :6.000
##     Heroine         Ketamine        LegalHighs         LSD
##  Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.000
##  Median :0.000   Median :0.0000   Median :0.000   Median :0.000
##  Mean   :0.374   Mean   :0.5692   Mean   :1.356   Mean   :1.062
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :6.000   Max.   :6.0000   Max.   :6.000   Max.   :6.000
##       Meth          MMushrooms       Nicotine        Semeron
##  Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000   1st Qu.:0.000000
##  Median :0.0000   Median :0.000   Median :3.000   Median :0.000000
##  Mean   :0.8265   Mean   :1.187   Mean   :3.201   Mean   :0.009549
##  3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:6.000   3rd Qu.:0.000000
##  Max.   :6.0000   Max.   :6.000   Max.   :6.000   Max.   :4.000000
##       VSA         Alcohol_User Amphetamine_User AmylNitrite_User Benzos_User
##  Min.   :0.0000   no :  68     no :1206         no :1515         no :1116
##  1st Qu.:0.0000   yes:1817     yes: 679         yes: 370         yes: 769
##  Median :0.0000
##  Mean   :0.4334
##  3rd Qu.:0.0000
##  Max.   :6.0000
##  Caffeine_User Cannabis_User Chocolate_User Cocaine_User Crack_User
##  no :  37      no : 620      no :  35       no :1198     no :1694
```

```
##  yes:1848      yes:1265      yes:1850       yes: 687     yes: 191
##
##
##
##
##  Ecstasy_User Heroine_User Ketamine_User LegalHighs_User LSD_User   Meth_User
##  no :1134     no :1673     no :1535      no :1123        no :1328   no :1468
##  yes: 751     yes: 212     yes: 350      yes: 762        yes: 557   yes: 417
##
##
##
##
##  MMushrooms_User Nicotine_User Semeron_User VSA_User
##  no :1191        no : 621      no :1879     no :1655
##  yes: 694        yes:1264      yes:   6     yes: 230
##
##
##
##
```

```
# Print the first observations of the data set
head(data_clean)
```

```
## # A tibble: 6 x 51
##      ID     Age Gender Education Country Ethnicity NScore EScore  OScore AScore
##   <dbl>   <dbl>  <dbl>     <dbl>   <dbl>     <dbl>  <dbl>  <dbl>   <dbl>  <dbl>
## 1     1  0.498   0.482   -0.0592   0.961     0.126  0.313 -0.575 -0.583  -0.917
## 2     2 -0.0785 -0.482    1.98     0.961    -0.317 -0.678  1.94   1.44    0.761
## 3     3  0.498  -0.482   -0.0592   0.961    -0.317 -0.467  0.805 -0.847  -1.62
## 4     4 -0.952   0.482    1.16     0.961    -0.317 -0.149 -0.806 -0.0193  0.590
## 5     5  0.498   0.482    1.98     0.961    -0.317  0.735 -1.63  -0.452  -0.302
## 6     6  2.59    0.482   -1.23     0.249    -0.317 -0.678 -0.300 -1.56    2.04
## # ... with 41 more variables: CScore <dbl>, Impulsiveness <dbl>,
## #   SensationSeeking <dbl>, Alcohol <dbl>, Amphetamine <dbl>,
## #   AmylNitrite <dbl>, Benzos <dbl>, Caffeine <dbl>, Cannabis <dbl>,
## #   Chocolate <dbl>, Cocaine <dbl>, Crack <dbl>, Ecstasy <dbl>, Heroine <dbl>,
## #   Ketamine <dbl>, LegalHighs <dbl>, LSD <dbl>, Meth <dbl>, MMushrooms <dbl>,
## #   Nicotine <dbl>, Semeron <dbl>, VSA <dbl>, Alcohol_User <fct>,
## #   Amphetamine_User <fct>, AmylNitrite_User <fct>, Benzos_User <fct>,
## #   Caffeine_User <fct>, Cannabis_User <fct>, Chocolate_User <fct>,
## #   Cocaine_User <fct>, Crack_User <fct>, Ecstasy_User <fct>,
## #   Heroine_User <fct>, Ketamine_User <fct>, LegalHighs_User <fct>,
## #   LSD_User <fct>, Meth_User <fct>, MMushrooms_User <fct>,
## #   Nicotine_User <fct>, Semeron_User <fct>, VSA_User <fct>
```

```r
# Evaluate if NAs are present
colSums(is.na(data_clean))
```

```
##               ID              Age           Gender        Education
##                0                0                0                0
##          Country        Ethnicity           NScore           EScore
##                0                0                0                0
##           OScore           AScore           CScore    Impulsiveness
##                0                0                0                0
## SensationSeeking          Alcohol      Amphetamine      AmylNitrite
##                0                0                0                0
##           Benzos         Caffeine         Cannabis        Chocolate
##                0                0                0                0
##          Cocaine            Crack          Ecstasy          Heroine
##                0                0                0                0
##         Ketamine       LegalHighs              LSD             Meth
##                0                0                0                0
##       MMushrooms         Nicotine          Semeron              VSA
##                0                0                0                0
##     Alcohol_User Amphetamine_User AmylNitrite_User      Benzos_User
##                0                0                0                0
##    Caffeine_User    Cannabis_User   Chocolate_User     Cocaine_User
##                0                0                0                0
##       Crack_User     Ecstasy_User     Heroine_User    Ketamine_User
##                0                0                0                0
##  LegalHighs_User         LSD_User        Meth_User  MMushrooms_User
##                0                0                0                0
##    Nicotine_User     Semeron_User         VSA_User
##                0                0                0
```

There are no *NA* values present in the data set, thus, no further pre-processing step is required.

Exploratory Data Analysis

In this section, an exploratory data analysis (EDA) is conducted in order to gain further insights about the Drug Consumption (Quantified) data set. These insights and characteristics of the data set will then be used to choose the appropriate models that are most suitable for the problem at hand as well as the three drugs used for user/non-user classification.

Used Features

For the classification of users and non-users, we will omit *country* and *ethnicity* in accordance to (Fehrman et al. 2015). Thus, ten input features remain:

- Age
- Gender
- Education
- N-Score (Neuroticism)
- E-Score (Extraversion)
- O-Score (Openness to Experience)
- A-Score (Agreeableness)
- C-Score (Conscientiousness)
- Impulsiveness
- Sensation-Seeking

Analysis Strategy

First, we will show the general distributions of age, gender and education of the whole data set in order to provide an overall impression of the participants.

Afterwards, we will select three drugs for evaluation based on the balance between users and non-users.

For each of the selected drugs, we will then show box plots of the four most discriminative features.

Age

```
# Plot histogram of age groups
data_clean %>%
  ggplot(aes(Age)) +
  geom_histogram()
```

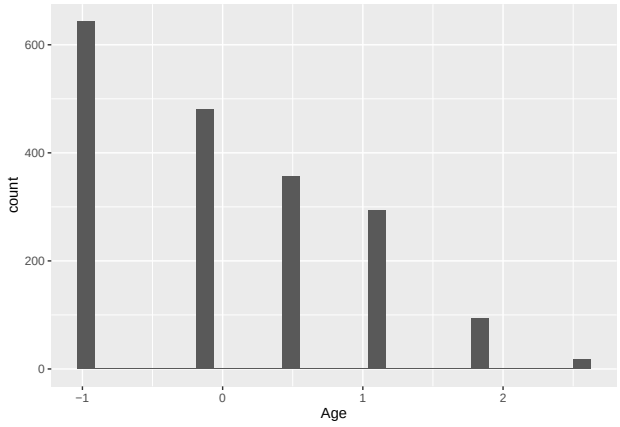


Figure 1: Histogram of 'Age'

Table 1: Mapping of numeric values of 'Age'

Value	Meaning
-0.95197	18 - 24
-0.07854	25 - 34
0.49788	35 - 44
1.09449	45 - 54
1.82213	55 - 64
2.59171	65+

Figure 1 shows the histogram of age groups in the data set. The mapping of numerical values to their corresponding meaning is shown in Table 1 and was taken from (Fehrman et al. 2015).

The age groups are not normally distributed. Instead, the distribution is skewed to the left, i.e., towards younger age groups.

Gender

```
# Plot histogram of gender
data_clean %>%
  ggplot(aes(Gender)) +
  geom_histogram()
```

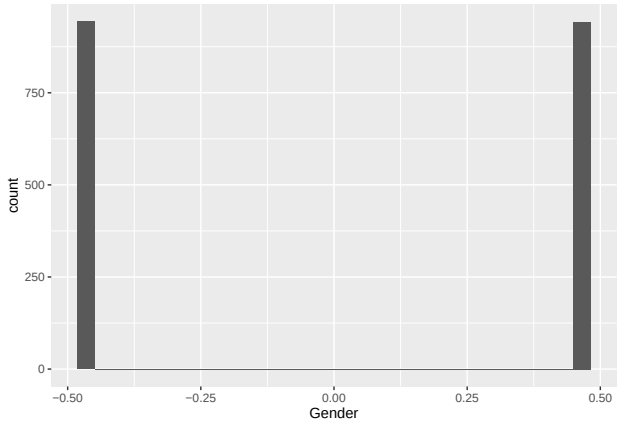


Figure 2: Histogram of 'Gender'

Table 2: Mapping of numeric values of "Gender"

Value	Meaning
-0.48246	Male
0.48246	Female

Figure 2 shows the histogram of genders in the data set. The mapping of numerical values to their corresponding meaning is shown in Table 2 and was taken from (Fehrman et al. 2015).

The genders "male" and "female" are uniformly distributed.

Education

```
# Plot histogram of education
data_clean %>%
  ggplot(aes(Education)) +
  geom_histogram()
```

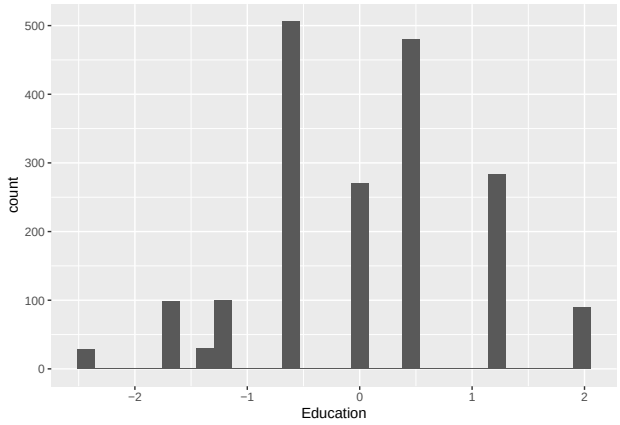


Figure 3: Histogram of 'Education'

Table 3: Mapping of numeric values of "Education"

Value	Meaning
-2.43591	Left school before 16 years
-1.73790	Left school at 16 years
-1.43719	Left school at 17 years
-1.22751	Left school at 18 years
-0.61113	Some college or university, no certificate or degree
-0.05921	Professional certificate / diploma
0.45468	University degree
1.16365	Masters degree
1.98437	Doctorate degree

Figure 3 shows the histogram of educations in the data set. The mapping of numerical values to their corresponding meaning is shown in Table 3 and was taken from (Fehrman et al. 2015).

The education is not regularly distributed. Most participants selected "Some college or university, no certificate or degree" (approx. 500 or 27%). Approximately 14% have left school at 18 years or younger. Another 14% have a professional certificate or diploma, which is also the average education. The remaining 45% have a university (25%), masters (15%) or doctorate degree (5%).

Class (User) Balance

Next, we examine the balance between users and non-users for each individual drug. We will select the top-3 drugs with the minimum difference between users and non-users.

```
# Compute difference between users and non-users for each drug
# and show the drugs with minimum difference
data.frame(yes = colSums(users_data),
           no = nrow(users_data) - colSums(users_data),
```

```
          drug = colnames(users_data)) %>%
  summarise(diff = abs(yes - no), drug = drug) %>%
  arrange(diff) %>%
  head()
```

```
##   diff             drug
## 1  347      Benzos_User
## 2  361  LegalHighs_User
## 3  383     Ecstasy_User
## 4  497  MMushrooms_User
## 5  511     Cocaine_User
## 6  527 Amphetamine_User
```

Accordingly, the drugs selected for evaluation in this report are:

- **Benzodiazepines**
- **Legal Highs**
- **Ecstasy**

Selected Drugs

For each of the selected drugs, we have compared differences between users and non-users for all ten input features (see Section 'Used Features') using box plots. For each of the selected drugs, we will show the four most discriminative features by means of the largest difference between the median of users and non-users, respectively.

Benzodiazepines

Regarding Benzodiazepines, an individual is more likely to be a user when the values for neuroticism, openness for experience, impulsiveness and sensation-seeking are greater than 0 (see Figure 4). As can be seen, the classification problem is rather hard, as none of the top-four features can exclusively separate users from non-users. For all input features, there is a large overlap between users and non-users.

Legal Highs

Regarding Legal Highs, an individual is more likely to be a user when the values for openness for experience and sensation-seeking are greater than 0 (see Figure 5). Additionally, there are stronger differences between users and non-users w.r.t. the age group and the education.

Ecstasy

Regarding Ecstasy, an individual is more likely to be a user when the values for openness for experience and sensation-seeking are greater than 0 (see Figure 6). As for Legal Highs, there are also stronger differences between users and non-users w.r.t. the age group and the education.

General Findings

For nearly all input features, there is a large overlap between users and non-users. For the three evaluated drugs, individuals with larger values for openness for experience and sensation-seeking are more likely to be a user.

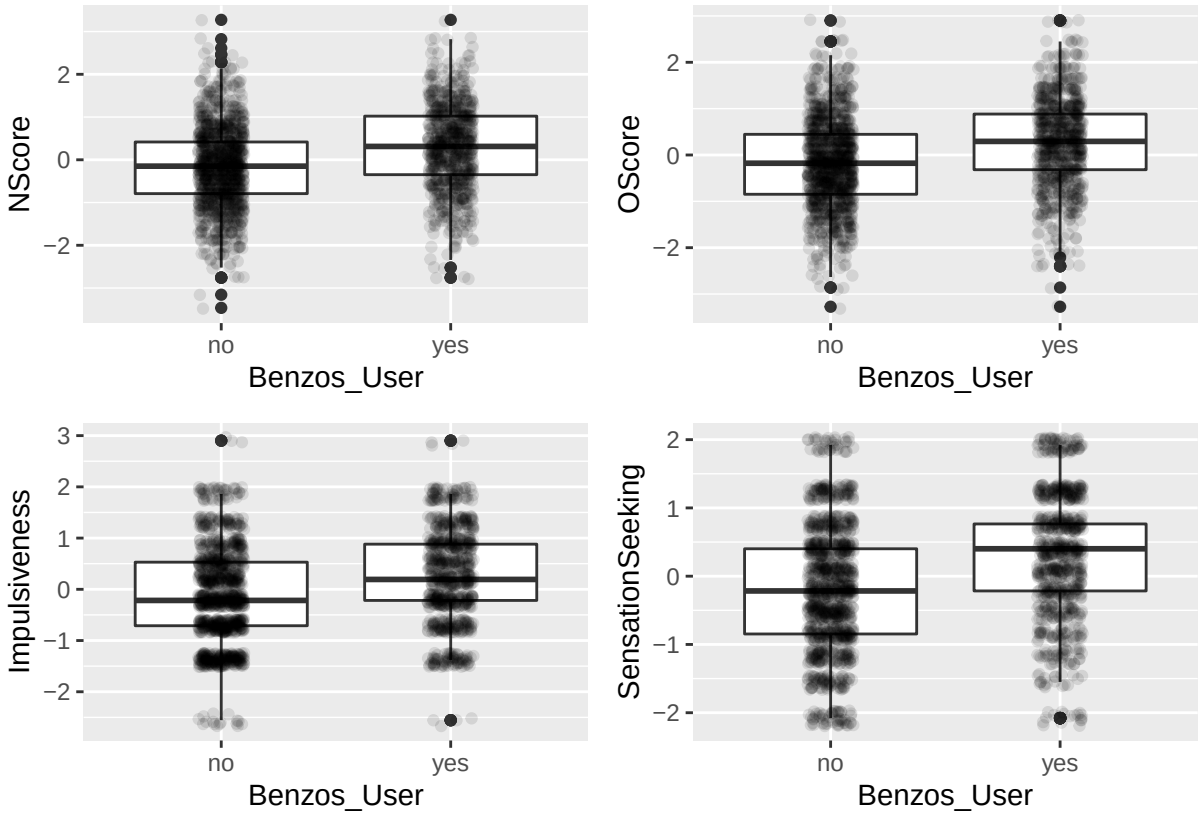


Figure 4: Four most discriminative features for Benzodiazepines

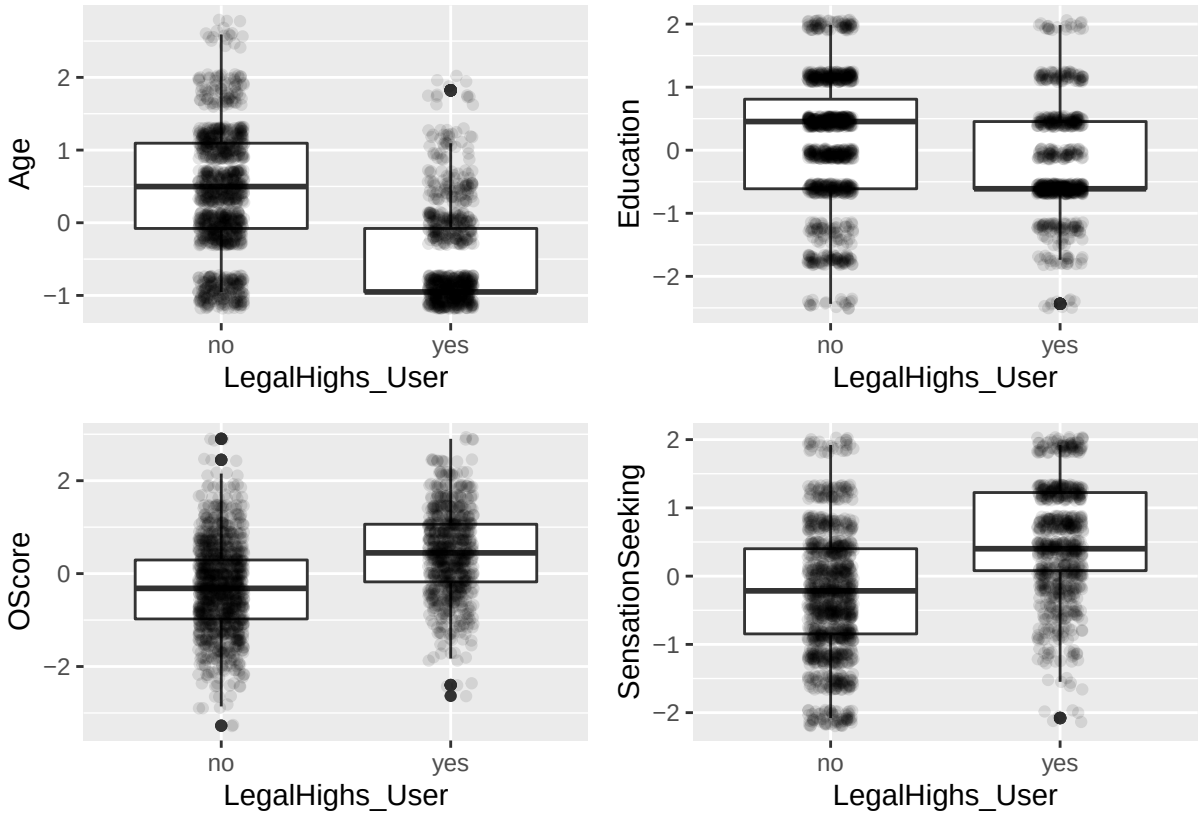


Figure 5: Four most discriminative features for Legal Highs

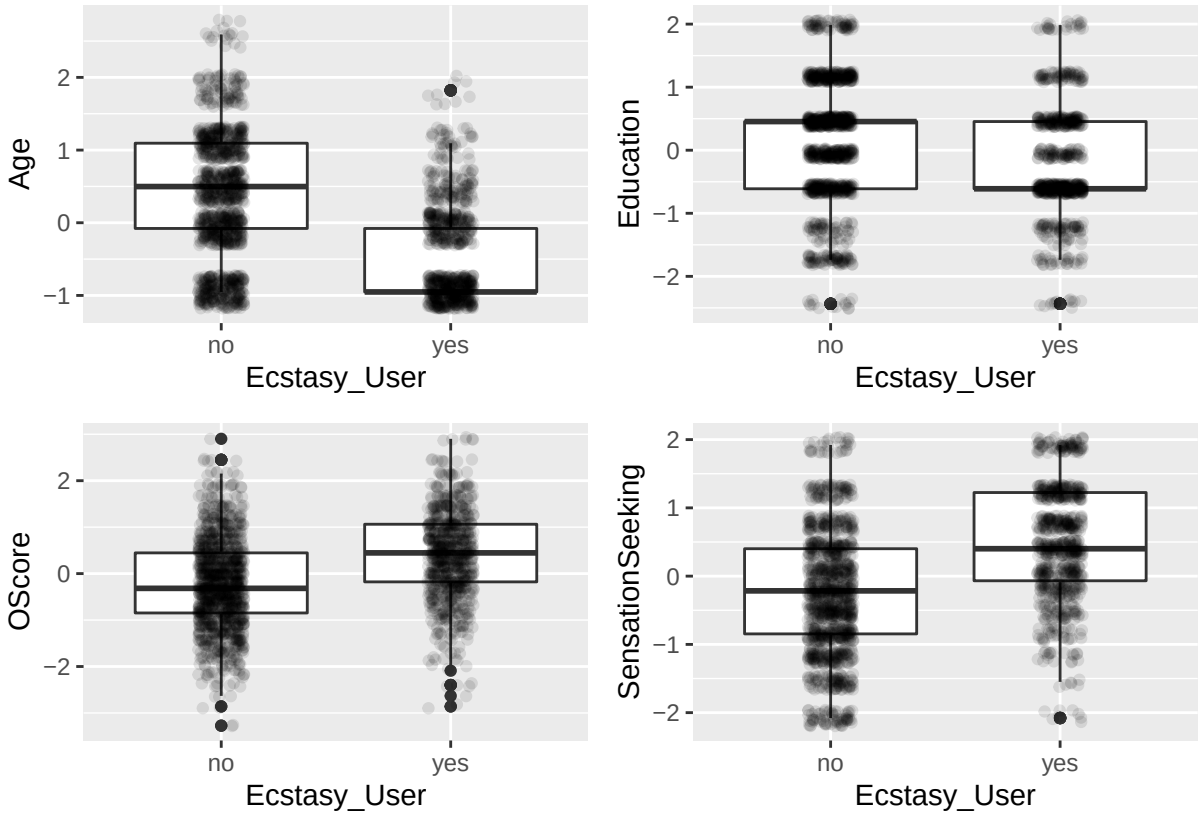


Figure 6: Four most discriminative features for Ecstasy

Methods

Based on the exploratory data analysis, we will use the three drugs Benzodiazepines, Legal Highs and Ecstasy for binary classification. We will use four different types of models that each try to separate users from non-users in ten-dimensional feature space:

- Boosted Generalized Linear Model (Boosted GLM)
- k-Nearest-Neighbor (kNN)
- Random Forest (RF)
- Regularized Discriminant Analysis (RDA)

These types of models were chosen because of their general applicability to binary classification problems. Boosted GLM and random forests are well-known ensemble methods that use boosting and bagging, respectively, as a powerful strategy to capture general feature-specifics from the training set.

If there are strong similarities between users of a specific drug, k-Nearest-Neighbor classification could be a fast and accurate approach. The data set is predestinated for distance metrics because of its quantification.

As a trade-off between linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), the regularized discriminant analysis (RDA) allows for quadratic decision boundaries in high-dimensional feature spaces.

The underlying theories are very briefly described in the following sub-sections. Please see the respective references for further details on these model types.

Boosted Generalized Linear Model

The Boosted Generalized Linear Model (Boosted GLM) was introduced and extended by (Buehlmann and Yu 2003) (Buehlmann 2006) (Buehlmann and Hothorn 2007). In principle, it uses the same approach as the well-known (generalized) linear models but in an ensemble fashion.

In Boosted GLM, simple component-wise – i.e., input-feature-wise – linear models are used as base learners. Then, gradient boosting is used to fit an ensemble model by optimizing a given loss function.

For our classification problem, we use the 0-1 loss function to maximize the classification accuracy. It takes the value 0 if the prediction matches the actual class, and the value 1 if not.

k-Nearest-Neighbor Classification

The algorithm k-nearest-neighbor was introduced by (Fix and Hodges 1951) and extended by (Altman 1992).

In this approach, no real training procedure is required. However, the parameter k should be optimized.

For each observation to be classified, the k-nearest neighbors in the given training set are determined using a distance measure. Thus, normalized training data is very beneficial for this approach. As explained in the quantification sub-section of the data set description, this is already the case for the used data set. Hence in this report, we simply use the Euclidean distance.

If the k-nearest neighbors are determined, the output class of the current test observation is given by the majority vote of these neighbors.

Random Forest

A random forest (RF) is an ensemble approach using simple decision trees as base learners. It was first created by (Ho 1995) and later extended by (Breiman 2001).

While training, usually several hundreds of trees are learned using bagging (bootstrap aggregating). This means that, for each tree, random samples are taken with replacement from the training set – usually a single or very few ones. Also, only a random subset of the available input features is used. This accounts for de-correlation of the trees.

Having trained the trees of the random forest, the classification result of a current test observation is given by the majority class vote.

Regularized Discriminant Analysis

The regularized discriminant analysis was proposed by (Friedman 1989) as a trade-off between the linear discriminant analysis (LDA) in the Fisher interpretation (Fisher 1939), where a common (pooled) covariance matrix estimate $\hat{\Sigma}$ is used, and the quadratic discriminant analysis (QDA) (Friedman and Tibshirani 2009), where an individual covariance matrix estimate $\hat{\Sigma}_k$ per class k is used. This allows for quadratic classification boundaries. The trade-off is controlled by two mixing factors λ and γ:

$$\hat{\Sigma}_k(\lambda, \gamma) = (1 - \gamma)\hat{\Sigma}_k(\lambda) + \gamma \frac{1}{p} \mathrm{tr}(\hat{\Sigma}_k(\lambda)) I \tag{2}$$

where p is the dimension of the measurement space. By optimizing λ and γ, the individual covariance matrix estimates $\hat{\Sigma}_k$ are shrunken towards a pooled covariance to a respective extent.

A discriminant (quadratic) score function is set up w.r.t. the distance of the current test observation and the k class means as well as the estimated covariance matrices $\hat{\Sigma}_k$ for all classes k.

To obtain the final classification, the current test observation is assigned to the class k for which the final score function is largest.

Results

In this section, we first describe the evaluation procedure and metrics. Afterwards, we briefly describe the results of (Fehrman et al. 2015) and (Mahyoub et al. 2019) that will be used for comparison. Finally, we present an overview of the results for the three evaluated drugs Benzodiazepines, Legal Highs and Ecstasy.

Evaluation Procedure

As proposed by (Mahyoub et al. 2019), we use 10-fold cross-validation on the complete data set to compute the results. More specifically, we use a 10-fold cross validation with a training/test-split of 90%/10% using a fixed seed – 2021 with sample kind "Rounding" – prior to every model generation to account for the repeatability of the experiments. We save the final predictions for each model and compare them to the actual observations to compute the evaluation metrics that will be described in the next sub-section.

Evaluation Metrics

For the final evaluation of the predictions, we will use the sensitivity, the specificity, the accuracy as well as the balanced accuracy. These metrics are defined as follows considering the confusion matrix:

Table 4: Confusion matrix

	True User	True Non-User
Predicted User	True Positive (TP)	False Positive (FP)
Predicted Non-User	False Negative (FN)	True Negative (TN)

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \tag{6}$$

Sensitivity and specificity were reported by (Fehrman et al. 2015) and used for choosing the optimal model and parameters per evaluated drug. (Mahyoub et al. 2019) reported accuracy, sensitivity and specificity. To account for imbalances of users and non-users, we also use the balanced accuracy which is the mean of sensitivity and specificity. We computed this metric for all comparison results.

Comparison to the State-of-the-art

We will use the results provided by (Fehrman et al. 2015) and (Mahyoub et al. 2019) for comparison.

(Fehrman et al. 2015) choose the best model performance by means of the maximum value of the minimum of sensitivity and specificity per drug. For the three evaluated drugs, a highly-optimized decision tree (DT) achieved the best performance. For each drug, 166 million decision tree models were tested optimizing for the splitting criterion, the set of input features, the minimum number of instances per leaf, and the weight of the class "users". The authors have used leave-one-out cross-validation (LOOCV) for computing the final results.

(Mahyoub et al. 2019) provides results of a decision tree (DT), a k-Nearest-Neighbor (kNN) classifier and a Random Forest (RF) per drug using 10-fold cross-validation.

Benzodiazepines

Table 5: Classification results for Benzodiazepines

Model	Sensitivity	Specificity	Accuracy	Balanced.Accuracy	Evaluation.Method
Boosted GLM	0.477	0.815	0.677	0.646	10-fold CV
kNN	0.495	0.779	0.663	0.637	10-fold CV
Random Forest	0.498	0.772	0.660	0.635	10-fold CV
RDA	0.553	0.777	0.685	0.665	10-fold CV
Mahyoub et al. DT	0.416	0.649	0.629	0.532	10-fold CV
Mahyoub et al. kNN	0.469	0.773	0.669	0.621	10-fold CV
Mahyoub et al. RF	0.521	0.749	0.675	0.635	10-fold CV
Fehrman et al. DT	0.709	0.715	NA	0.712	LOOCV

Legal Highs

Table 6: Classification results for Legal Highs

Model	Sensitivity	Specificity	Accuracy	Balanced.Accuracy	Evaluation.Method
Boosted GLM	0.730	0.850	0.802	0.790	10-fold CV
kNN	0.699	0.815	0.768	0.757	10-fold CV
Random Forest	0.699	0.843	0.785	0.771	10-fold CV
RDA	0.739	0.844	0.802	0.792	10-fold CV
Mahyoub et al. DT	0.576	0.711	0.705	0.643	10-fold CV
Mahyoub et al. kNN	0.649	0.803	0.764	0.726	10-fold CV
Mahyoub et al. RF	0.581	0.747	0.756	0.664	10-fold CV
Fehrman et al. DT	0.795	0.824	NA	0.809	LOOCV

Ecstasy

Table 7: Classification results for Ecstasy

Model	Sensitivity	Specificity	Accuracy	Balanced.Accuracy	Evaluation.Method
Boosted GLM	0.655	0.809	0.747	0.732	10-fold CV
kNN	0.670	0.772	0.731	0.721	10-fold CV
Random Forest	0.664	0.791	0.741	0.728	10-fold CV
RDA	0.647	0.817	0.750	0.732	10-fold CV
Mahyoub et al. DT	0.532	0.653	0.644	0.593	10-fold CV
Mahyoub et al. kNN	0.589	0.765	0.703	0.677	10-fold CV
Mahyoub et al. RF	0.594	0.711	0.709	0.652	10-fold CV
Fehrman et al. DT	0.762	0.772	NA	0.767	LOOCV

Discussion

For the three evaluated drugs, our regularized discriminant analysis (RDA) model achieved the best performance in 10-fold cross-validation out of the four tested model types – slightly better than Boosted GLM. Compared to (Mahyoub et al. 2019), our RDA model outperforms their presented results by means of balanced accuracy:

- 66.5% (RDA) vs. 63.5% (RF of (Mahyoub et al. 2019)) for Benzodiazepines
- 79.2% (RDA) vs. 72.6% (KNN of (Mahyoub et al. 2019)) for Legal Highs
- 73.2% (RDA) vs. 67.7% (KNN of (Mahyoub et al. 2019)) for Ecstasy

Of course, these results are subject to variation w.r.t to the randomness in training/test-splits within each fold of the cross-validation. (Mahyoub et al. 2019) didn't provide the seed used for their experiments (if applicable). We made sure to use a fixed seed – 2021 with sample kind "Rounding" – for all our experiments to account for repeatability.

However, our RDA results are not as good as the results achieved with a highly-optimized decision tree (DT) by the authors of the data set (Fehrman et al. 2015):

- 66.5% (RDA) vs. 71.2% (DT of (Fehrman et al. 2015)) for Benzodiazepines
- 79.2% (RDA) vs. 80.9% (DT of (Fehrman et al. 2015)) for Legal Highs
- 73.2% (RDA) vs. 76.7% (DT of (Fehrman et al. 2015)) for Ecstasy

In this comparison, there is a first difference in the evaluation method – 10-fold cross-validation vs. leave-one-out cross-validation (LOOCV). In LOOCV, the randomness of training/test-splits is eliminated. For this data set, there are 1,885 folds (one for each observation) each with 1,884 training samples and 1 test sample. Thus, there is much more training data available for each fold. On the other hand, the runtime needed for completing the experiment is increased by two orders of magnitude.

The next difference is related to optimization. Inter alia, (Fehrman et al. 2015) optimized for the set of input features ending up with the following sets for the three evaluated drugs:

- **Benzodiazepines**: Age, Gender, N-Score, E-Score, Impulsiveness, Sensation-Seeking
- **Legal Highs**: Age, Gender, O-Score, A-Score, C-Score, Sensation-Seeking
- **Ecstasy**: Age, Gender, Sensation-Seeking

As can be seen above, four to seven input features were discarded in the optimal decision trees for the three evaluated drugs. Thus, using the complete set of input features for our RDA models might be sub-optimal as well.

Our Boosted GLM and RDA models achieved the highest specificities (i.e., true negative rates) of all reported results.

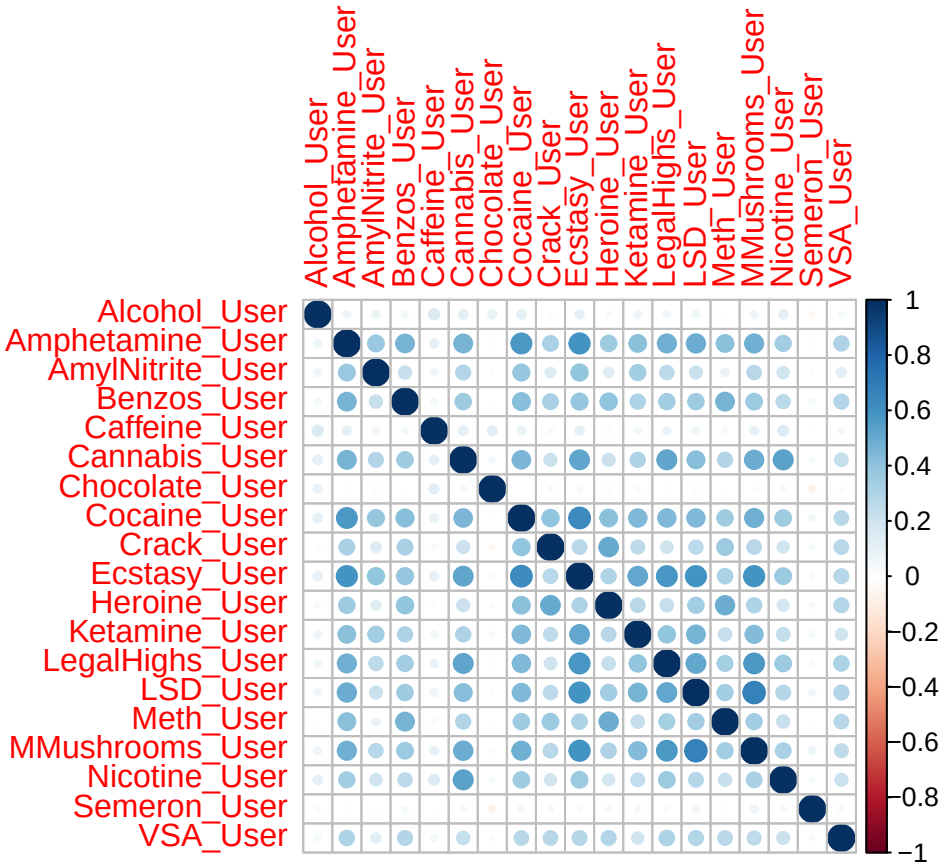
There is still room for improvement regarding our best model performance. In future work, we aim at performing a more detailed optimization of model parameters and the set of input features. Additionally, we will use LOOCV to increase comparability between our results and the ones of (Fehrman et al. 2015). Instead of discarding input features, we would like to present a first experiment on adding other input features in the following sub-section.

Correlation Between Drugs

As presented by (Fehrman et al. 2015), there can be a strong correlation between the user-groups of different drugs. It is also often observed in the data set that individuals are users of multiple drugs. Hence, we perform a small experiment to estimate whether the classification performance increases if we would know that the individual is a user or non-user of a correlated drug.

First, we select the top-2 correlated drugs for each of the evaluated drugs:

```
corrplot(cor(users_data))
```



According to this output, we will use the following additional input features:

- **Benzodiazepines**: Amphetamine_User, Meth_User
- **Legal Highs**: Ecstasy_User + MMushrooms_User
- **Ecstasy**: Cocaine_User + MMushrooms_User

The achieved results are presented in the following tables:

Benzodiazepines

Table 8: Experimental classification results for Benzodiazepines

Model	Sensitivity	Specificity	Accuracy	Balanced.Accuracy	Evaluation.Method
RDA	0.553	0.777	0.685	0.665	10-fold CV
RDA with user info	0.648	0.847	0.766	0.747	10-fold CV
Fehrman et al. DT	0.709	0.715	NA	0.712	LOOCV

Legal Highs

Table 9: Experimental classification results for Legal Highs

Model	Sensitivity	Specificity	Accuracy	Balanced.Accuracy	Evaluation.Method
RDA	0.739	0.844	0.802	0.792	10-fold CV
RDA with user info	0.867	0.825	0.842	0.846	10-fold CV
Fehrman et al. DT	0.795	0.824	NA	0.809	LOOCV

Ecstasy

Table 10: Experimental classification results for Ecstasy

Model	Sensitivity	Specificity	Accuracy	Balanced.Accuracy	Evaluation.Method
RDA	0.647	0.817	0.750	0.732	10-fold CV
RDA with user info	0.872	0.832	0.848	0.852	10-fold CV
Fehrman et al. DT	0.762	0.772	NA	0.767	LOOCV

These results show that knowing whether an individual is a also a user or non-user of a highly correlated drug drastically improves the classification results, in particular by increasing the specificity (true positive rate).

Model Insights

Insights about RDA models are hardly to visualize from the resulting model object, i.e., which input features were the most important and where are the decision boundaries. Nevertheless, our boosted generalized linear models achieved nearly similar classification performance. We now show the variable importance of the trained models to reveal potential input features that might be discarded in future work:

Benzodiazepines

```
## glmboost variable importance
##
##                  Overall
## NScore           100.000
## OScore            88.953
## SensationSeeking  71.419
## Gender            56.272
## Education         32.056
## CScore            25.689
## AScore            19.569
## Impulsiveness      8.122
## EScore             6.449
## Age                0.000
```

Legal Highs

```
## glmboost variable importance
##
##                  Overall
## Gender            100.00
## Age                99.68
## OScore             56.21
## SensationSeeking   55.18
## EScore             19.96
## CScore             17.93
## Education          11.28
## AScore              0.00
## NScore              0.00
## Impulsiveness       0.00
```

Ecstasy

```
## glmboost variable importance
```

```
##
##                   Overall
## Age                100.00
## SensationSeeking    82.49
## Gender              65.86
## OScore              43.06
## CScore              24.37
## Education           17.01
## EScore               0.00
## Impulsiveness        0.00
## NScore               0.00
## AScore               0.00
```

It can be seen that the manually selected four most discriminant features have a non-zero variable importance for each drug, respectively. In fact, most of them are among the most important variables. The variable importance outputs show that three to four variables per drug were not able to support the classification into users and non-users. This is a starting point for further optimization in future work.

Conclusion

In this report, we have performed binary classification of users and non-users for the three drugs Benzodiazepines, Legal Highs and Ecstasy on the **Drug Consumption (Quantified)** data set. We have used four types of classification models per drug: Boosted generalized linear models, k-nearest-neighbor, random forests and regularized discriminant analysis (RDA). Our results were also compared to the results reported by (Fehrman et al. 2015) and (Mahyoub et al. 2019), respectively.

Using the same input features, we achieved a better performance than (Mahyoub et al. 2019) for the three evaluated drugs w.r.t. to all evaluation metrics. The performance of our best model (RDA) was slightly worse than the highly optimized decision trees of (Fehrman et al. 2015), which also determined the optimal feature selection for each drug.

Overall, we have achieved balanced accuracies of 66.5%, 79.2% and 73.2% using RDA for the drugs Benzodiazepines, Legal Highs and Ecstasy, respectively.

As an outlook, we provided first experimental results when using additional input features, namely whether the individual is also a user or non-user of correlated drugs, and have shown a significantly improved performance.

Further improvements are expected by using an optimal input feature sub-set for each drug as well as a more sophisticated tune grid.

References

Altman, Naomi S. 1992. "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression." *The American Statistician* 46 (3): 175–85.

Beaglehole, R., R. Horton, C. Adams, G. Alleyne, P. Asaria, V. Baugh, H. Bekedam, N. Billo, and S. Casswell. 2011. "Priority Actions for the Non-Communicable Disease Crisis." *The Lancet* 377 (9775): 1438–47.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Buehlmann, Peter. 2006. "Boosting for High-Dimensional Linear Models." *The Annals of Statistics* 34 (2): 559–83.

Buehlmann, Peter, and Torsten Hothorn. 2007. "Boosting Algorithms - Regularization, Prediction and Model Fitting." *Statistical Science* 22 (4): 477–505.

Buehlmann, Peter, and Bin Yu. 2003. "Boosting with the L2 Loss - Regression and Classification." *Journal of the American Statistical Association* 98 (462): 324–39.

Costa, P. T., and R. R. MacCrae. 1992. *Revised NEO-Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI). Professional Manual. Odessa, FL, Psychological Assessment Resources.*

"Drug Consumption (Quantified) at UCI Machine Learning Repository." 2016. 2016. https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29.

Fehrman, E., A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban. 2015. "The Five Factor Model of Personality and Evaluation of Drug Consumption Risk." *Data Science.* https://arxiv.org/abs/1506.06297.

Fisher, Ronald A. 1939. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7 (2): 179–88.

Fix, Evelyn, and Joseph L. Hodges. 1951. *Discriminatory Analysis - Nonparametric Discrimination - Consistency Properties. USAF School of Aviation Medicine, Randolph Field, Texas.*

Friedman, Jerome Hastie, and Robert Tibshirani. 2009. *The Elements of Statistical Learning, Volume 2. Springer Series in Statistics, New York, NY, USA.*

Friedman, J. H. 1989. "Regularized Discriminant Analysis." *Journal of the American Statistical Association* 84 (405): 165–75.

Ho, Tin Kam. 1995. "Random Decision Forests." *Proceedings of the 3rd International Conference on Document Analysis and Recognition* 1.

Kleiman, M. A., J. P. Caulkins, and A Hawken. 2011. *Drugs and Drug Policy - What Everyone Needs to Know. Oxford University Press.*

Lee, S. Y., W. Y. Poon, and P. M. Bentler. 1995. "A Two-Stage Estimation of Structural Equation Models with Continuous and Polytomous Variables." *British Journal of Mathematical and Statistical Psychology* 48 (2): 339–58.

Linting, M., and A. van der Kooij. 2012. "Nonlinear Principal Component Analysis with CatPCA - A Tutorial." *Journal of Personality Assessment* 94 (1): 12–25.

Mahyoub, Mohammed A., Alaith Abu Lekham, Emad Alenany, Lubna Tarawneh, and Daehan Won. 2019. "Analysis of Drug Consumption Data Using Data Mining Techniques and a Predictive Model Using Multi-Label Classification." *Proceedings of the IISE Annual Conference.*

Martinson, E. O., and M. A. Hamdan. 1971. "Maximum Likelihood and Some Other Asymptotically Efficient Estimators of Correlation in Two Way Contingency Tables." *Journal of Statistical Computation and Simulation* 1 (1): 45–54.

Stanford, M. S., C. W. Mathias, D. M. Dougherty, S. L. Lake, N. E. Anderson, and J. H. Patton. 2009. "Fifty Years of the Barratt Impulsiveness Scale – an Update and Review." *Personality and Individual Differences* 47 (5): 385–95.

Zuckerman, M. 1994. *Behavioral Expressions and Biosocial Bases of Sensation Seeking. Cambridge University Press.*