STATS 503 Group 31 Project
Predicting Risk of Drug Misuse

Zhijun Hua, Jingyan Lu, Yuying Wang, Siwei Tang

April 27, 2020

1 Introduction

Drug abuse is a precarious behavior that can lead to poor health for individuals and potentially cause social disorder[1]. Therefore, it is critical to model an individual's potential drug misuse risk and find solutions to eliminate detrimental attempts. Multiple factors – societal, demographic, and psychological, etc. – are believed to shed lights on drug use behaviors. Apart from commonly used factors (i.e. age, gender and education), personality traits have been proposed to impact on drug abuse rate. In this project, the Five Factor Model (FFM) is adopted to explain personality traits. FFM is generally considered as a comprehensive taxonomy for personality traits in terms of five dimensions: Neuroticism(N), Extraversion(E), Openness to Experience(O), Agreeableness(A), and Conscientiousness(C).

The primary goal of this study is to evaluate an individual's risk of drug misuse based on personality characteristics (FFM along with another two important factors: impulsiveness and sensation seeking), and general demographic information.

2 Data Set

2.1 Data Description

The data set `Drug consumption (quantified) Data Set` [2] is obtained from UCI Machine Learning Repository with 1885 instances and 32 attributes which include an index variable, 12 predictors, and frequency of use of 19 legal and illegal drugs. The original data was collected from an anonymous online survey seeking participants at least 18 years of age mainly from English-speaking countries: the UK, the USA, Canada, Australia, New Zealand and Ireland. The data available to us are already quantified, so all 12 input features can be viewed as real-valued and they are listed below:

1. `Age`: 6 groups
2. `Gender`: 2 groups
3. `Education`: 9 groups
4. `Country`: 7 groups
5. `Ethnicity`: 7 groups
6. `N-score`: Neuroticism
7. `E-score`: Extraversion
8. `O-score`: Openness
9. `A-score`: Agreeableness
10. `C-score`: Conscientiousness
11. `Impulsive`: Impulsiveness
12. `SS`: Sensation seeing

4 specific drug types: `Cannabis, Heroin, Magic Mushrooms, and Nicotine` are chosen in this project. Among these 4 drugs, Nicotine is legal, Heroin is illegal, while Cannabis and Mushrooms are not clearly defined because of ambiguity in legal provisions. We believe this combination of legal and illegal, commonly used and seen drugs falls under general interest. Furthermore, it allows us to compare different effects of predictors on drug use.

2.2 Data Preprocessing

2.2.1 Data Cleaning

The data set is loaded into R for data analysis. All 1885 instances have no missing values and all 12 predictors have been quantified, so no further data cleaning steps are necessary.

2.2.2 Recoding Variables

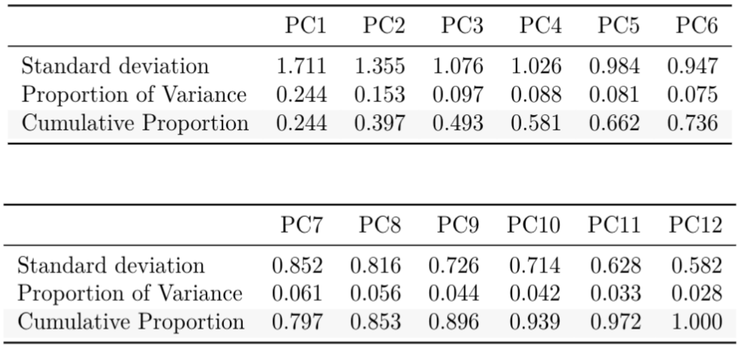
The response variable `frequency of use` for each drug type was categorized into 7 groups. We recoded them into binary cases for a clearer definition of drug users and non-users, and sufficient sample size for each group. The new binary classes are shown below:

1. `User`: Used in
    - Last Decade
    - Last Year
    - Last Month
    - Last Week
    - Last Day
2. `Non-User`
    - Never Used
    - Used over a Decade Ago

Now the processed data contains 12 predictors and 4 response variables. Out of total 1885 instances, 70 percent are split into training set, and the rest 30 percent are treated as testing data.

2.2.3 Dimension Reduction

Principal Component Analysis is implemented on scaled training set to examine the possibility of reducing the number of predictors and eliminating unnecessary ones, while preserving as much explanatory power as possible. As shown in the results below, to account for 90 percent of the total variance at least 9 principal components are required. Thus, we regard it unworthy to sacrifice model's interpretability for such a small amount of reduction in dimension. The following steps are proceeded without dimension reduction.



	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.711	1.355	1.076	1.026	0.984	0.947
Proportion of Variance	0.244	0.153	0.097	0.088	0.081	0.075
Cumulative Proportion	0.244	0.397	0.493	0.581	0.662	0.736

	PC7	PC8	PC9	PC10	PC11	PC12
Standard deviation	0.852	0.816	0.726	0.714	0.628	0.582
Proportion of Variance	0.061	0.056	0.044	0.042	0.033	0.028
Cumulative Proportion	0.797	0.853	0.896	0.939	0.972	1.000

Table 1: PCA Summary Table

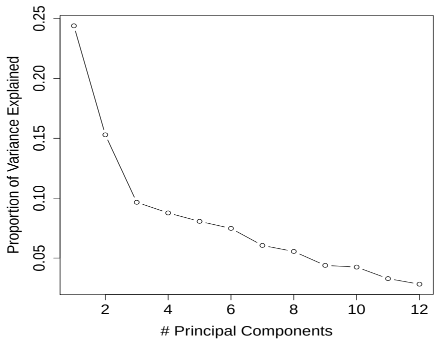


Figure 1: Scree Plot

3 Exploratory Data Analysis

3.1 Summary Statistics

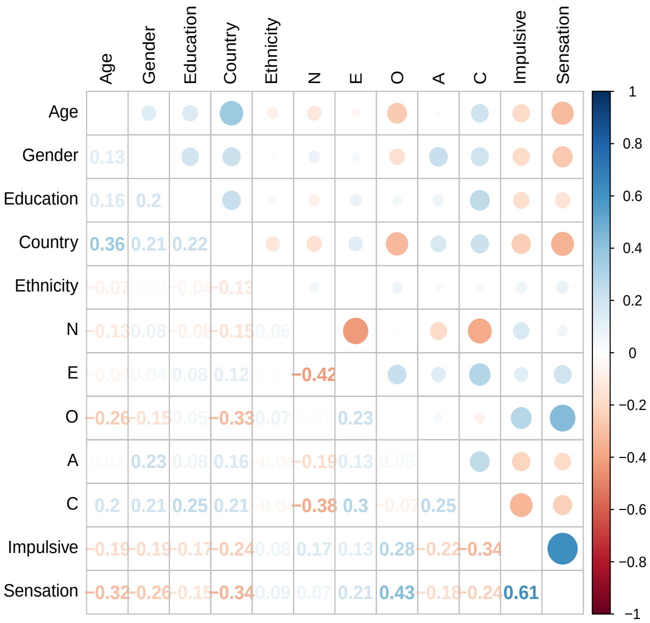
The training data are composed of 1319 instances in total. The distribution of two classes for each drug type is shown below. As indicated by the table, sample sizes in each category are not well balanced due to the new definition applied here and its dependence on how dangerous each type of drug is. However, the overall data set is satisfactory, and its sufficient sample size will give us reliable model selections.

	Cannabis	Heroin	Mushrooms	Nicotine
`Non-User`	447	1166	854	431
`User`	872	153	465	888

Table 2: Distribution of Classes for 4 Classification Problems

3.2 Correlation between Predictors

Since all 12 predictors are real-valued, we check their correlations directly from Figure 2 below. It is obvious that most pairs are not heavily or even moderately correlated. The only pair that really stands out is between **Impulsive** and **Sensation** with correlation coefficient 0.61. Thus, it will not be surprising to find only one of them significant in the following analysis. Again, the weakly correlated predictors support our previous choice of not applying dimension reduction here.



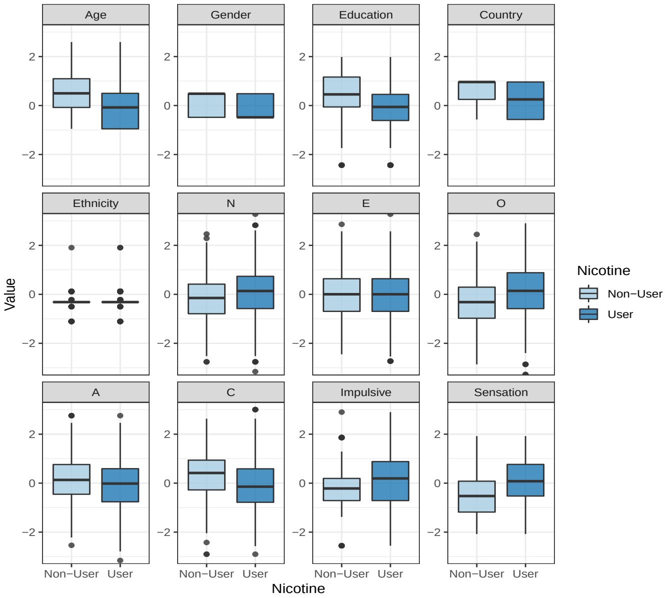


Figure 2: Correlation Plot

Figure 3: Boxplots Grouped by Classes

3.3 Boxplot

From Figure 3 above, we notice that Age, Education, Country, N, O, C, Impulsive and Sensation seem to be useful in identifying 2 classes for drug Nicotine. Although not shown here due to space limit, same graphs are plotted for the other three classification problems and we find a similar set of variables to be useful. To summarise some interesting patterns, users of all four drug types seem to be at a younger age, more open to new experience, lack of conscientiousness and more impulsive. Country has a varying effect on different drug use.

3.4 Density of Predictors

The distributions of each predictor are shown in Figure 4 below. The 5 personality traits (N, E, O, A, C) seem to follow a somewhat normal distribution. The demographic variables hardly follow any distribution. Even though they have real values, each possible value corresponds to a category in the original data.

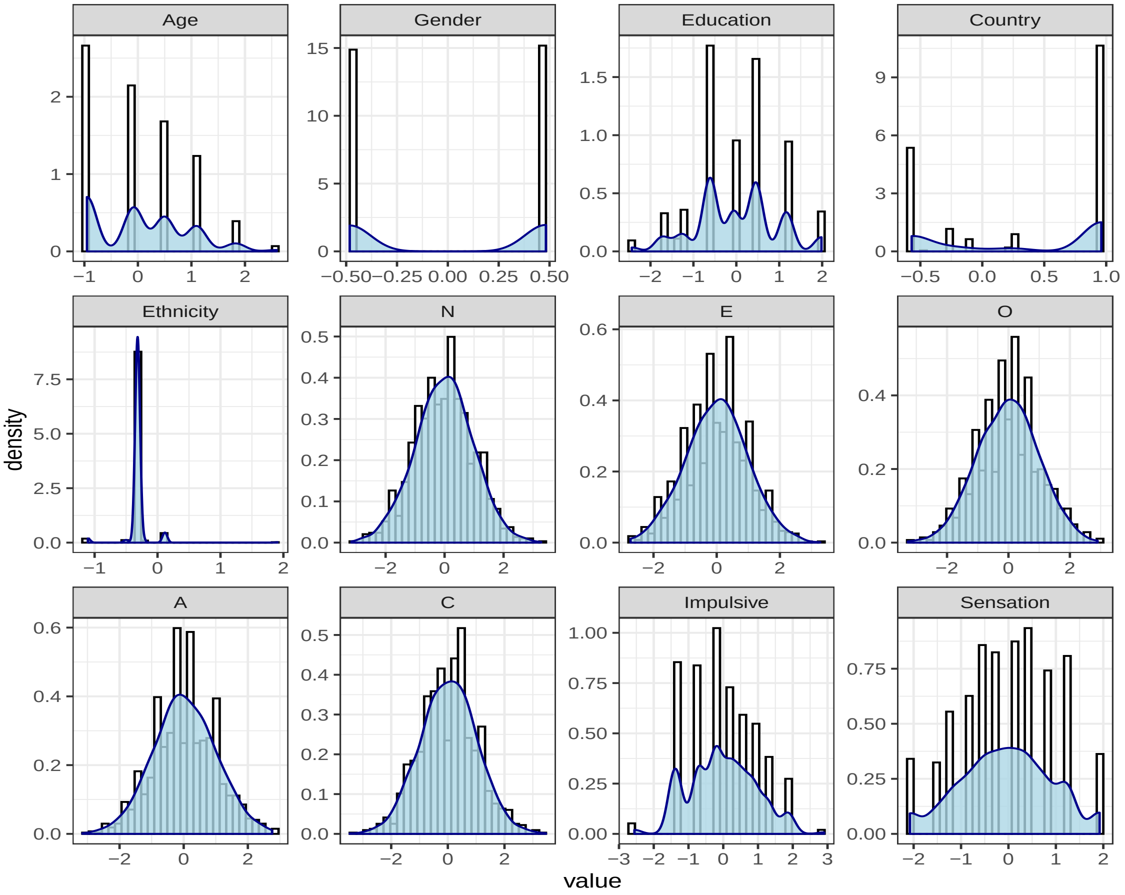


Figure 4: Density Plot

4 Methods

4.1 Parametric Classification

4.1.1 Linear Discriminant Analysis(LDA)/Quadratic Discriminant Analysis(QDA)

We apply both linear and quadratic discriminant analysis to the dataset. In both methods, it is assumed all the distributions for observations from different classes are normal distributions. The unknown parameters are the mean μ_k and covariances Σ_k of normal distribution for the kth class, as well as the prior probability π_k of a random sample from the class k. These parameters are estimated from the training data and then plugged into the discriminant functions[3].

The difference between LDA and QDA is that we assume all the covariances are equal in LDA, which leads to linear boundary between classes. However, the assumption does not hold for QDA and the boundary of QDA is quadratic.

The discriminant function for LDA is

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + log\pi_k.$$

And the discriminant function of QDA is

$$\delta_k(x) = -\frac{1}{2}log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + log\pi_k.$$

4.1.2 Logistic Regression

We use the standard logistic regression here by fitting the response variable using all 12 explanatory variables. Moreover, we predict drug misuse by

$$P(\text{Drug Use} = \textbf{User}|X) > 0.5.$$

4.1.3 Naive Bayes

We implement the standard Naive Bayes on our data. Naive Bayes method assume Gaussian distributions for numeric variables. The prior are calculated from the proportion of the training data.

4.2 Non-parametric Classification

4.2.1 KNN

K-nearest neighbour classifies the object as the majority class of its k nearest neighbours. Euclidean distance is used in this project.

We apply KNN method separately on the 4 drug types. 10-fold Cross Validation is used to look for the best-K for each drug. Then we use the best-K's for each drug to fit the data.

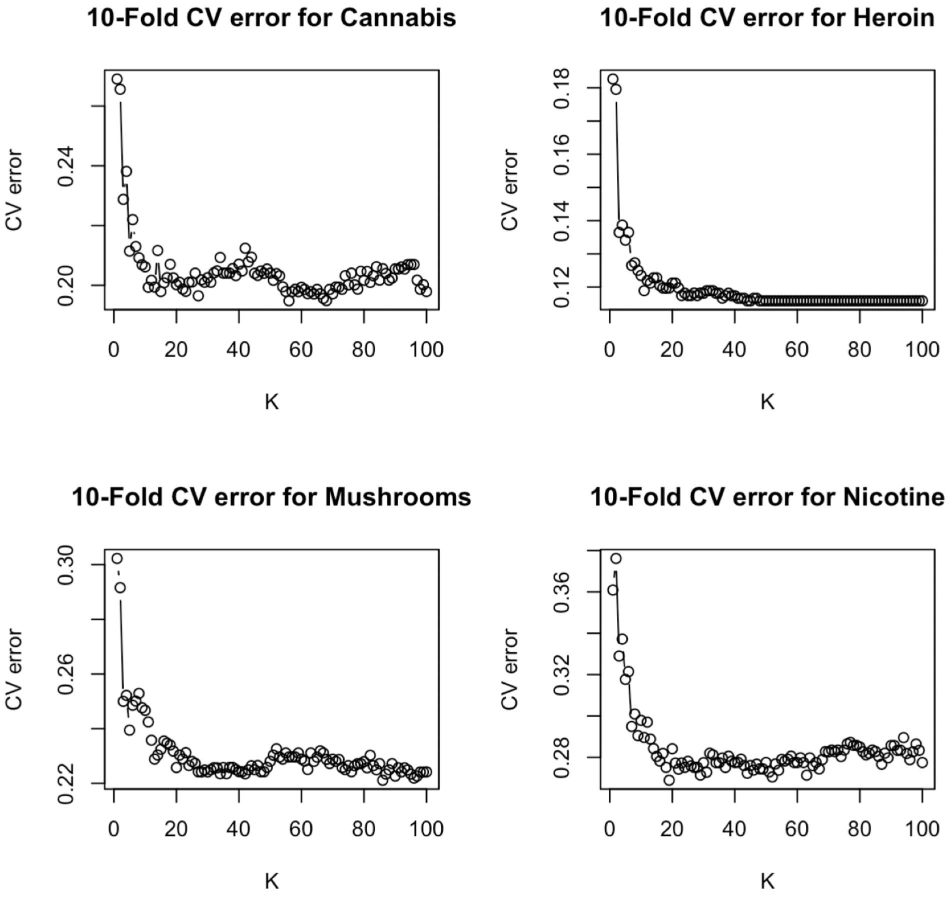


Figure 5: 10-fold Cross Validation

The best-K's chosen by 10-fold Cross Validation for all four response variables are shown in the table below.

	Cannabis	Heroin	Mushrooms	Nicotine
Best K	56	44	86	19

Table 3: Best K for 4 Different Drug Types

4.2.2 Decision Tree/Random Forest/Bagging/Boosting

Decision Tree
In this study, all methodologies within the purview of tree classification have been tested. The decision tree is evaluated as the best performed model among all tree methods with regard to the test error. In this project, we are targeting at growing the decision tree with its pruning parameter *Cost-Complexity Pruning*, which aims to penalize the number of trees in the model. The model is formulated as the following.

$$\sum_{1}^{|T|} N_t \times Impurity_t + C_p \times |T| \tag{1}$$

The Cp plots shown below demonstrate how we select *Cp* parameters based on the relative errors. We apply Gini coefficient as the default splitting rules.

$$Gini(node_m) = \sum_{k=1}^{K} p_k(m) \times (1 - p_k(m)) \tag{2}$$

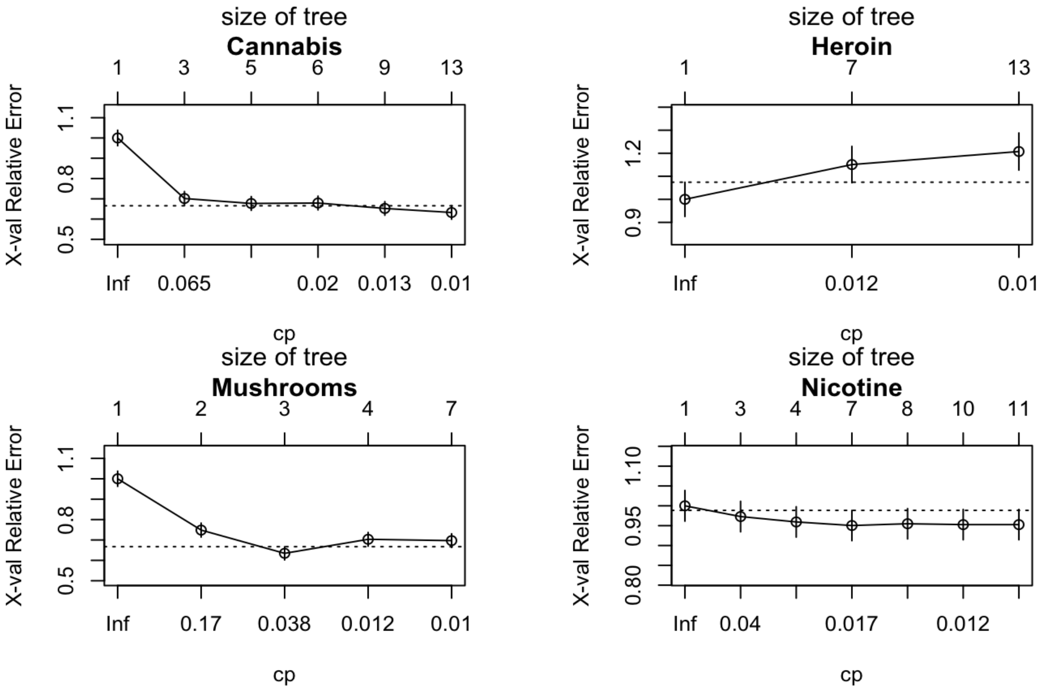


Figure 6: *Cp* Plot for Cannabis, Heroin, Mushrooms and Nicotine

Due to model simplicity's concern, the smallest *Cp* values are not preferred when relative errors are close. In view of the four plots(Figure 6), we select `0.01` for `Cannabis`, `0.012` for `Heroin`, `0.038` for `Mushroom`, and `0.017` for `Nicotine`.

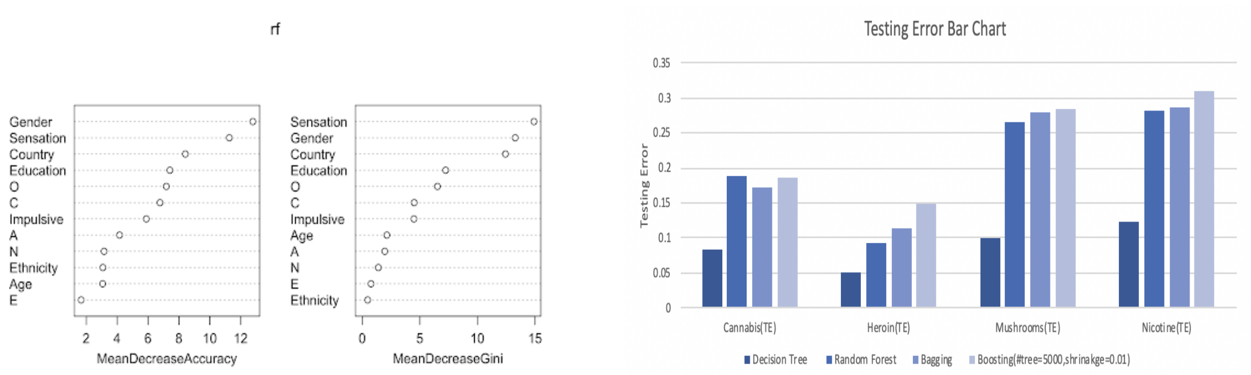
	Cannabis	Heroin	Mushrooms	Nicotine
`Decision Tree`	0.08649469	0.04779970	0.10015175	0.12063733

Table 4: Testing Error of Decision Tree

Bagging and Random Forest

Bagging is an ensemble classification approach that fits many large trees to bootstrap re-sampled version of the training data, and classifies by the majority vote. Random Forest builds on the idea of bagging, and provides an improvement by de-correlating the tree through random selection on variables. A random subset of m predictors is chosen as split candidates from the full set of p predictors. By default, we set

$$m \approx \sqrt{p} \tag{3}$$



(a) Variable Importance in Random Forest

(b) Tree Method Test Error Bar Chart

Figure 7

Graph (a) in Figure 7 shows that the three most important variables through Random Forest is `Gender`, `Sensation`, and `Country`. Graph (b) on the right demonstrates the testing error(TE) for all four tree models, among which the decision tree has the smallest testing error. By default in Bagging, we set number of trees to be 1000 and m=p=12; and in Random Forest we found that the $m = \sqrt{p}$, where p is equal to 12.

	Cannabis	Heroin	Mushrooms	Nicotine
`Bagging`	0.17314488	0.11484099	0.27915194	0.28621908
`Random Forest`	0.18727915	0.09363958	0.26501767	0.28268551

Table 5: Testing Error of Bagging and Random Forest

Adaboosting

$$f(x) = \sum_{m=1}^{M} \mu\alpha_m T_m(x) \tag{4}$$

Adaboost builds an additive model by stagewise fitting the model using the exponential loss function. In this study, we found that setting `shrinkage`=0.01 instead of default 0.1 slightly improves the testing error, and increasing number of trees from 1000 to 5000 does enhance the performance as well. Therefore, we set the parameter `shrinkage` to be 0.01 and `ntree` to be 5000.

	Cannabis	Heroin	Mushrooms	Nicotine
`Adaboost`	0.18551237	0.14840989	0.2844523	0.30918728

Table 6: Test Error of Adaboosting

5 Results

We apply all the methods shown above on the training data set and use the model to predict the testing set. Test errors are shown in Table 7.

	Cannabis	Heroin	Mushrooms	Nicotine
`LDA`	0.205	0.117	0.235	0.291
`QDA`	0.233	0.154	0.215	0.265
`Logistic Regression`	0.182	0.102	0.242	0.253
`Naive Bayes`	0.171	0.217	0.265	0.260
`KNN`	0.175	0.104	0.270	0.295
`Decision Tree`	**0.086**	**0.048**	**0.100**	**0.121**

Table 7: Test Error of Different Methods

Among all the methods, `decision tree` has the `smallest` test error for all 4 drug types. A majority of the methods find smaller test errors for `Heroin` and `Cannabis` than for `Mushrooms` and `Nicotine`. Since `Heroin` and `Cannabis` are illegal drugs, we believe that illegal drug use can be better predicted.

For our best model decision tree, we reach the same conclusion that it has better performance when predicting usage on illegal drugs `Heroin` and `Cannabis`. Taking a closer look at the best tree, we find that Decision Tree selects 9 variables for `Heroin` and 7 variables for `Cannabis`, while the other 2 models only contain 2 and 4 variables. Therefore, it is not surprising to find larger test errors for legal drug use given small tree sizes. Noticing that decision tree generates test errors all below 0.121. This decent result shows no major sign of overfitting.

In general, we find people at younger age who are open to experience and seeking for feelings more likely to be classified as `Users` for legal drug. Cannabis Users from counties outside the UK tend to be younger or more open to experience while Cannabis Users from the UK seek for more feelings and are relatively lack of conscientiousness. Heroin Users tend to be from the USA or other countries with lower levels of education. These findings approximately correspond to our exploratory data analysis.

6 Conclusion

Our project focuses on the `Drug consumption (quantified) Data Set`. The purpose is to classify individuals into binary classes (User and Non-User) according to demographic information and personal traits for 4 drugs: `Cannabis, Heroin, Mushrooms` and `Nicotine`. Throughout the project, a total set of nine classification methods have been implemented, and `decision tree` has the best performance with the lowest test error. Moreover, illegal drugs Heroin and Cannabis tend to have lower test error than the other drugs.

In the future, we could extend our project to try more complex classification methods such as Support Vector Machines and Neural Networks. Besides, we can use this dataset to solve multi-class classification problem and check if it provides better prediction error. One improvement could be, since drug users have various personalities depending on different countries, having more samples from each country will give us even better results on determining the possible factors on drug use.

References

[1] Fehrman, Elaine Muhammad, Awaz Mirkes, Evgeny Egan, Vincent. (2017). *The Five Factor Model of Personality and Evaluation of Drug Consumption Risk.*

[2] Drug Consumption Dataset. Retrieved from `http://archive.ics.uci.edu/ml/machine-learning-databases/00373/`.

[3] James, Gareth Witten, Daniela Hastie, Trevor Tibshirani, Robert. (2013). *An Introduction to Statistical Learning: With Applications in R.*