Predicting Drug Use, Legal and Illegal

DePaul DSC 540-Winter 2019

Kyle Spanski

**Abstract**

This paper is an attempt to find what makes a person use certain drugs. The data come from an online survey of 1,885 respondents regarding 18 different drugs. I first found that the usage of some drugs seemed to be uncorrelated with others, mainly the commonly used legal drugs in the study (Alcohol, Caffeine, and Chocolate). The target variables were unbalanced when looking at the 7 classes of use, so I updated the targets to a binary value (respondents have ever used the drug or never used the drug). Personality traits affect use of all drugs strongly, though the traits differ for the various drugs. In fact, the top two or three factors that affect usage the most vary widely and often include demographic features. Various approaches were tested and it appears that a Random Forest approach with Feature Selection (wrapper method) for use in the last year yields the most accurate and parsimonious model.

**Introduction**

For many these days there is a focus on the harm that certain drugs cause and what can be done to prevent or slow usage. It seems there is always a new drug in the news that scares a certain percentage of the population. In this paper, I will explore the reasons people use certain drugs. With this information, hopefully more can be done to prevent certain susceptible individuals from using these harmful drugs in the first place or at least notifying those who are susceptible. I will compare both very harmful and illegal drugs and less harmful drugs to see what characteristics these users share and how they differ.

In this paper, I generally treat the drugs as separate classification problems. There is some analysis of combinations of drugs but the main focus is the classification of each drug on its own.

**Literature Review**

One study found that people take different drugs for different reasons [1]. This could mean that we see different contribution factors for different drugs in our analysis. A meta study [2] found that there are two distinct categories of young drinkers (10-25 years old), those that drink for enhancement motives and those for coping motives. This bimodality of personality types is something to look for in those who use other drugs.

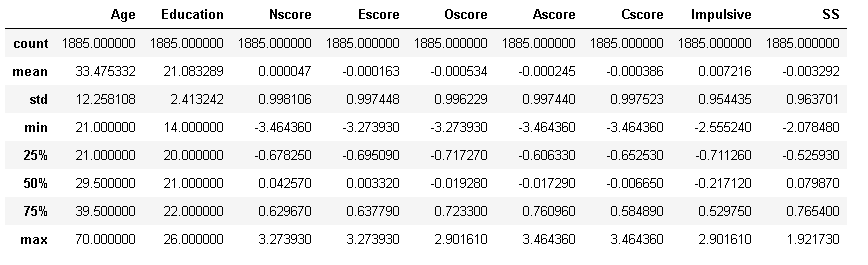
One study of adolescents in Iran found that the majority of users had traumatic childhood experiences and came from dysfunctional families [3]. This could be another indicator that personality and past experience contribute to drug use. The study that generated the data set used in this paper [4] mentions there are certain “Pleiades” or clusters of drugs that are commonly used together, a phenomenon also investigated in this paper. Finally, a study in 1990 [5] found that certain psychological characteristics were significant predictors in whether teens used marijuana, again suggesting personality is a factor in drug use.

**Methodology**

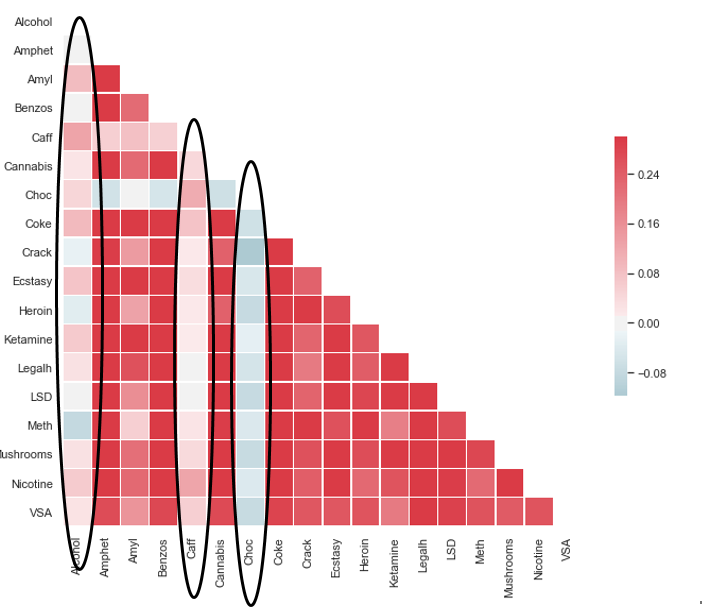
The data is made up of 1885 instances (respondents) with 31 variables for each instance. There are 7 different personality attributes (like Neuroticism, Agreeableness, Impulsivity etc.) all based on questionnaire answers. There is also demographic data including Gender, Age Group, Education, Country and Ethnicity. The target variables are the usage scores for 18 different drugs (17 real and 1 made up to correct for over-claimers).

All analysis was done in Jupyter Notebook using the Python programming language. The SKLearn and Pandas packages were used extensively. All of the features in the data were provided as numeric values and the target values were strings.

The personality attributes were normalized floats so I left those as is. The Age Group variable was converted to the median of that age group for easier calculation. I also updated the Education variable to the number of years of education represented (or median where appropriate). I made the Gender, Ethnicity, and Country variables into strings which I then converted to dummy variables for each feature value. After this conversion, the numeric features are summarized below:

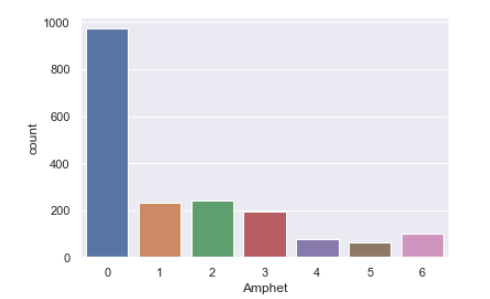


Next, I then dropped any observation where the fake “Semer” drug was marked as being used and dropped that “Semer” target variable (dropped 8 observations). I then updated the string target variables (CL0, CL1, CL2 etc.) to their numeric equivalents. I ran a correlation heat map using the Pearson Correlation Coefficient as seen below:

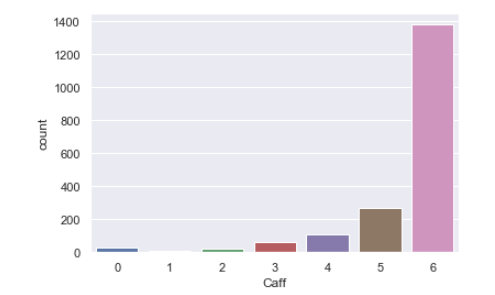


We can observe that the three legal drugs (Alcohol, Caffeine, and Chocolate) do not seem to be correlated with the other illegal drugs. The illegal drugs all seem to have correlated use with each other.

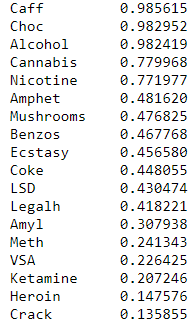
The next step was to look at the usage histograms of individual drugs. The use of most of the illegal drugs was very unbalanced with most respondents never having used the drug and the users spread out between the other classes. Amphetamine is a good example of this as seen below:



The legal drugs seemed to be skewed in the other direction with users generally using these every day or at least weekly. Caffeine is an example below:



To account for this imbalance, I reset all the target variables to just two classes based on whether the respondent had ever reported using the drug. This provided a much more balanced set of target variables as seen in the table of the mean of the target classes (percent of respondents who have ever used the drug) below:



Next I looked at the Variance Inflation Scores for all of the variables and found that the newly created dummy variables had VIF scores that were too high in the model (infinity). I corrected for this by dropping one of the values for each dummy variables and this brought the scores back to a reasonable level of under 10 for each explanatory variable.

With all of the features and targets converted, I was ready to run Machine Learning Models for all of the drugs individuals to predict whether a respondent had ever used the drug in question. The first model was a basic Decision Tree with no Feature Selection. This would provide a good baseline to compare against both more complex models as well as the percentage of users for that drug. I’m most interested in how much better the model does at predicting a user than random chance (percentage of users for that drug). I was able to try different Cross-Validation folds and settled on 5 being the most ideal in terms of both runtime and spread of results.

Next I ran Decision Trees with Feature Selection turned on. I used the Wrapper method which is looking to see whether each feature has more importance than the mean of the other features. If it does, that feature is kept in the model.

To get a more illustrative picture of what features are important to different drugs, I created and visualized Decision Trees for each of the drugs but limited the tree depth to just two levels (along with Feature Selection).

Next I looked at Random Forests. Again I used Feature Selection as it did not have a significant effect on performance and removed around half the features. I also ran a Random Forest model on a 70/20/10 Train/Test/Evaluation split of the data to see whether the data was balanced. I was looking for similar performance on the Test and Evaluation data sets.

After the original Random Forest models, I ran the Support Vector Classifier models to see if they proved to be better classifiers. I tried two kernel types, RBF and Linear.

I then used the original numeric target classes (0-6) to run a regression model to predict which specific user class the respondent would be in instead of just whether they have ever used the drug.

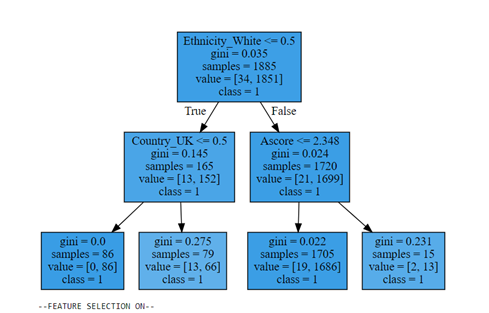
Finally I wanted to look at a different time frame for use and transformed the target variable into a binary “Used in Last Year” variable. I ran Random Forests again for both all of the original drugs and a combined “CokeCrack” engineered target where the respondent was coded as a user if they used either drug or both in the last year.

**Results**

I will present the results of just one drug (Amphetamines) with all of the model approaches described above. These results are a good representation of the other illegal drugs with fairly even target value classes (Amyl Nitrite, Benzodiazepine, Cocaine, Ecstasy, LSD, Methamphetamine and Magic Mushrooms).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Amphetamines Results** | Accuracy | Area Under the Curve | RMSE | Explained Variance |
| Decision Tree-No Feature Selection (CV=5) | 0.60 (+/- 0.05) | 0.60 (+/- 0.05) | N/A | N/A |
| Decision Tree - With Feature Selection (CV=5) (10/22 features selected) | 0.58 (+/- 0.05) | 0.58 (+/- 0.04) | N/A | N/A |
| Decision Tree - With Feature Selection and Max Depth = 2 (CV=5) | 0.63 (+/- 0.12) | 0.70 (+/- 0.11) | N/A | N/A |
| Random Forest - With Feature Selection and Max Depth = 5 (8/22 features selected) | 0.68 (+/- 0.09) | 0.74 (+/- 0.08) | N/A | N/A |
| Random Forest with Train / Test / Eval (70/20/10) Split (Scores for 10% Evaluation Set) (No CV) | .665 | .736 | N/A | N/A |
| Support Vector Classifier (Kernel = RBF) | 0.66 (+/- 0.09) | 0.75 (+/- 0.06) | N/A | N/A |
| Support Vector Classifier (Kernel = Linear) | 0.64 (+/- 0.14) | 0.74 (+/- 0.09) | N/A | N/A |
| Random Forest Regression | N/A | N/A | 1.55 (+/- 0.81) | 0.09 (+/- 0.24) |
| Random Forest (Use in Past Year) - With Feature Selection and Max Depth = 5 (8/22 features selected) | 0.77 (+/- 0.06) | 0.83 (+/- 0.13) | N/A | N/A |

Another interesting view is seen when I looked at the 2-level Decision Trees, example for Alcohol below:



Comparing these trees, we can see that the most important few features differ greatly with the drug we’re looking at. Demographics (Ethnicity, British, and the Agreeableness) are the most important features for Alcohol prediction and about half the drugs. In other drugs (Amphetamines, Amyl Nitrite, Chocolate, Cocaine, Ecstasy, Ketamine and Nicotine) the most important feature is a personality measure. When not limiting the number of features, every iteration included five of the personality measures (Narcissism, Extraversion, and Openness to Experience, Agreeableness, and Conscientiousness).

The new “CokeCrack” combined target performed about as well in a Random Forest model as the regular Cocaine drug did which was the lower performer of the two. There did not seem to be any lift in model performance when combining these two drugs.

**Discussion**

In summary, the Random Forest Classifier was the most accurate classifier for this data set. Regardless of what timeframe is chosen, Random Forest with Feature Selection is both the most accurate and parsimonious. The Support Vector Classifier is not significantly worse performing, but I chose Random Forest as it is easier to explain. Since we’re dealing with some still unbalanced target variables, it’s important to measure performance against the Area Under the Curve generally more than the Accuracy especially for the very unbalanced legal drugs.

The five personality traits from the NEO-FFI-R Personality Inventory were important features and kept in every model. It’s important to look at lower depth trees for each of the drugs as the most important few features vary greatly.

**Future Work**

I would like to have more localized data in this study. There is a lot of variance in the people in different areas of the US and UK so City or State data would be interesting. I would also like to compare the results from the online survey with use data from the general population. This would probably increase the bias of the targets but would be a worthwhile analysis. I could also explore even more groupings of drugs with more than just two drugs at a time and use different timeframes (used in last week, used in last day etc.)

**Conclusion**

This dataset shows that personality measures explain a meaningful percentage of drug use. However, it is important to look at drugs individually as the most important driver of use differs by the drug. It is also necessary to think about the timeframes and patterns of drug use when doing any analysis as some drugs are used more or less frequently than others. If I was tasked with predicting new person’s drug use, I would feed their data into a Random Forest Classifier with Feature Selection to both predict accurately and be able to explain the model to any interested parties.

**References**

1-Boys, Annabel, et al. “What Influences Young People’s Use of Drugs? A Qualitative Study of Decision-Making.” Drugs: Education, Prevention & Policy, vol. 6, no. 3, Nov. 1999. EBSCOhost, doi:10.1080/09687639997052.

2-Kuntsche, Emmanuel, et al. “Who Drinks and Why? A Review of Socio-Demographic, Personality, and Contextual Issues Behind the Drinking Motives in Young People.” Addictive Behaviors, vol. 31, no. 10, 2006, pp. 1844–1857., doi:10.1016/j.addbeh.2005.12.028.

3-MIRLASHARI, JILA, et al. “Early Childhood Experiences, Parenting and the Process of Drug Dependency among Young People in Tehran, Iran.” Drug & Alcohol Review, vol. 31, no. 4, June 2012, pp. 461–468. EBSCOhost, doi:10.1111/j.1465-3362.2011.00384.x.

4-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 [https://arxiv.org/abs/1506.06297], 2015

5-Shedler, J., & Block, J. (1990). Adolescent drug use and psychological health: A longitudinal inquiry. American Psychologist, 45(5), 612-630.

http://dx.doi.org/10.1037/0003-066X.45.5.612