Drug Usage Prediction from Personality Traits

Brandon McIntyre

mcint170@msu.edu

Final Report

Introduction

Monitoring drug usage of the population is useful for understanding the overall health, and risks certain populations have. Most of the drugs looked at in this study will be that of the illegal kind with varying effects on overall health. Some drugs like heroin are concerning due to the high dependency and negative health effects it has on the user. Being that these drugs are dangerous and illegal it is worth investigating ways we can predict drug usage. If it is known personality trait contribute to drug use this can help target the populations that might need help the most to understand the risks taking such drugs can have on their overall health and life. This information can also be useful to understand if there is certain traits that are more likely to lead to drug use, which may help guide therapy for those that are users of this drug and are looking to get help. This project aims to use only personality traits and some demographic information to determine if it is possible to predict current drug usage based on personality traits. This project will utilize linear and quadratic discriminant analysis, random and boosted trees, and support-vector-machines. The hope is to create models that can accurately predict, given personality trait scores and some demographics, if a patient has used a certain drug within the past decade.

Related Work

The work that was specifically looked at prior to this study was looking at current drug usage patterns in "Baby Boomers" that were aged 50 to 64 years old [1]. This work was primarily focused on summary statistics from a SAMHDA study. Unfortunately, the link to that original 2012 data is unavailable, but an update dataset for a similar study in 2016 is located at SAMHDA's website[1]. Also provided by the article is their own custom dataset that uses part of the 2012 study[2] . This custom dataset can be used to re-create the articles analysis.

In an attempt to understand the analysis further, I recreated the bar plot of the original article from the custom dataset. This can be seen on the left hand side of Figure 1. I also attempted to recreate the bar plot using a new dataset from UCI[3]. I will speak more about this dataset in the next section, but this data comes from an online survey conducted from 2011 to 2012. The bar plot created from the UCI data will be on the right hand side of Figure 1.

---

[1]https://www.datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2016-nsduh-2016-ds0001-nid17185

[2]https://github.com/fivethirtyeight/data/tree/master/drug-use-by-age

[3]https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29

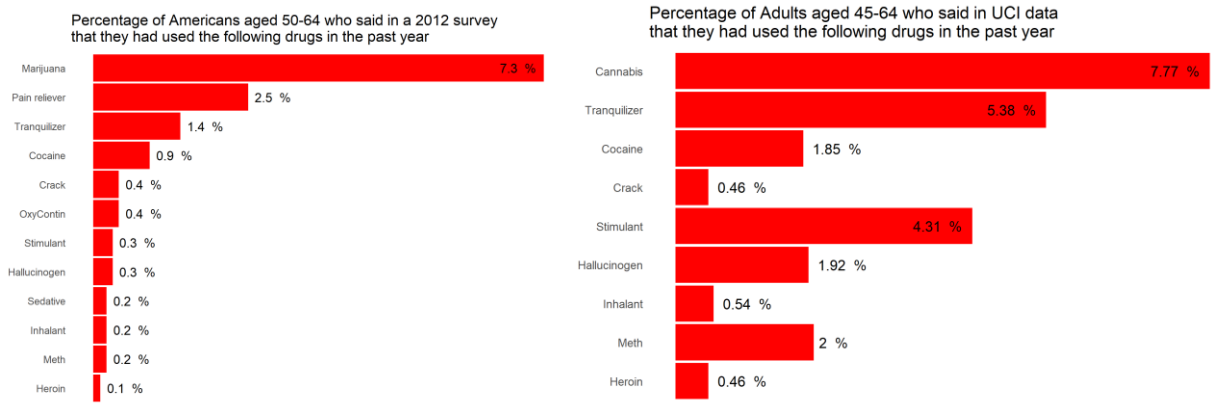

Figure 1: Drug usage of Adults between ages 45 and 64. The left panel is the recreation of the original plot from the 538 Article. The right panel is creating a similar bar plot using UCI data. Note that not all sections were possible to recreate because of different available data, so the UCI data has less categories. Also, the UCI plot is using any adult between age 45 and 64, whereas the 538 plot is using only Americans between 50 and 64. Also it could be possible different drugs fall under different categories as it was not clear in 538 study which drug feel under the broader categories such as stimulant.

There are some differences between the two bar plots. This is possible due to many factors such as the fact the plot with UCI data is using all Adults, the age groups between the plots are not exactly the same, or that the drugs may be counted different towards different border categories . Also, obviously, they come from different participants as the 538 data comes from SAMHDA and the UCI Data comes from an online survey with most likely different sample sizes. However, despite coming from different data there is still some interesting similarities in the data. Cannabis was the most used, and Heroin was the least used. Cocaine was used more than crack for both as well. It is also interesting to find that even though there are differences they both seem to have values quite near each other with values beneath 8%.

Dataset

The dataset used for the rest of the project is the UCI dataset. The dataset comes from a 2012 study from Fehrman, E. [2] and was altered and normalize to the form it is now by UCI. The data consists of demographic information on age, gender, education, and country of residence for each participant. In addition to that 6 personality attributes were recorded for each respondent regarding to neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. Also recorded was participants use of 18 different drugs/substances: alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine, volatile substance abuse, and one fictitious drug called “Semeron” to catch “over-claimers”.

The data dealing with drug use measures “Never Used”, “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”. The data for the entire dataset has been transformed and normalized. This makes the data actually easier to use as we will not have to use factors in our models. This also allows things like QDA to work because it alleviates issues with linear dependence between rows of data.

For this study, the data dealing with drug usage will be transformed into binary to allow for our classification modeling. If the participant used that particular drug within the last decade we will mark them as a user. If last used a decade ago they will be marked as non-user. These binary labels

is what we will use to train and create prediction models of if certain personality traits lead to using one drug or another. See Figure 2 for the percentage of people that have used this drug within the last decade.

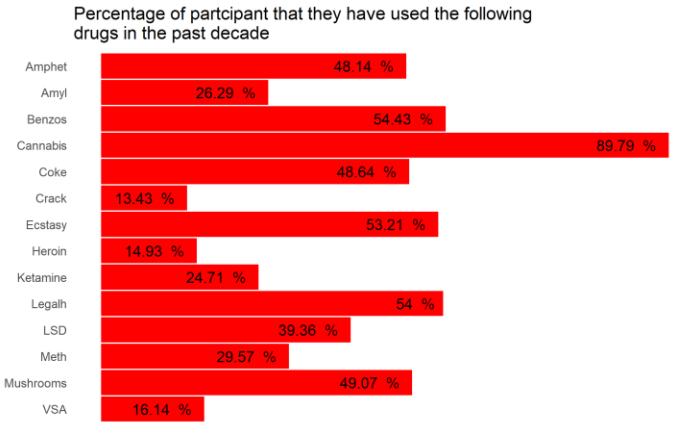

Figure 2. This bar plot shows the percentage of participants that have used each drug within the last decade.

As we can see some Figure 2, we at least will have 13% of the data corresponding to positive matches for every drug. Some are more common than others. We will see if these differences in abundance effect model performance.

Method

The complete derivation of the models used will be skipped in this report as this can be lengthy. Instead, I will briefly describe the model and how it its accuracy was obtained. Each drug was processed the same, and for each model an accuracy was obtained.

The accuracy for every model is simply the mean of if the model guessed correctly. Meaning if all the predictions were correct the mean would be 1. This is because if the prediction is correct it corresponds to 1 and, if not, it will be 0. Thus, the sum of this divided by the total observations will give us our accuracy.

Linear Discriminant Analysis - LDA

For the first model, I used LDA. LDA is a classic method for classification. It uses the idea that the data has gaussian distribution and each variable class has the same variance. Using this idea, it is able to find a linear discriminating line between classes.

To find the accuracy of the model, I used Leave One Out Cross Validation (LOOCV). This simply uses all but 1 data point to train. Then it fits that one remaining data point for the test. It repeats this process until all data points have been tested on. The average of the predictions is then recorded and this is what is used for the prediction that we get our accuracy measurement from

In order to implement LDA with LOOCV, I simply used the function *lda* from the package *MASS* with the setting *CV=TRUE*

Quadratic Discriminant Analysis - QDA

This is model very similar to LDA and the way LDA works. However, instead of assuming each class has the same variance, we now consider that each class has its own separate variance. This also means now that our discriminating line between classes will become quadratic.

Finding the accuracy with LOOCV is the same as LDA. Instead, we will just use the *qda* function from *MASS* and use the same argument of *CV=TRUE*

Random Forest – RF

This model uses the idea of decision trees in order to predict a value. The decision trees in random forest, however, are constructed in a special way. For each fork in a tree, only a random subsection of variables can be chosen. This introduces variability in the tree building process. Then many trees are produced this way. Then each tree is used to predict the value, and either the average of all tree's output are used, or majority vote is taken and the most occurring outcome is chosen.

With this method to find the accuracy I simply used a validation set for training and testing data. This is because I did not vary any parameters within the model, but simply used the default. There was not a need to use multi-fold cross validation.

The model was implemented using the function *randomForest* from the package *randomForest*. The model used default setting for number of variables to subset and the depth to create to.

Boosted Tree

As with the random forest method this also uses a decision tree. However, instead of creating a bunch of random large trees. The boosted tree focuses on creating small trees. The small trees take into account the reduction in MSE between each fork. It keep the trees small because it has a penalty term for making too big of trees.

This model I used a slightly different approach. Ideally, we would want to use k-fold cross validation because for this model I am testing for the best shrinkage value. However, this model takes time to run, and due to time constraints, I was only able to use a k-fold of 1. This means for every different shrinkage value, I tested with a training set. I think found the shrinkage variable that gave me the best accuracy and I used that model for the final test set. This was the accuracy value I ended up using.

This model was implemented using the *gbm* function from the package *gbm.* I varied the shrinkage value from 0.003 to 1.47. I also used 1000 trees for the number of trees to use.

Principal Component Analysis -PCA.

Before using the Support-Vector-Machine, I decided to clean up my data a little bit with PCA. What PCA does for us is allows us to the find the "Principal Component Vectors", otherwise known as eigenvectors, of our data. With this we are able to clean our data because after finding these Principal components we can deconstruct the data and reconstruct it only using the principal components. This acts as a way to "clean our data" and only retain the information that will help us classify our values with Support-Vector-Machine.

To preform PCA I used the function *prcomp* which comes from the package *stats* which is a basic package of R. I used default settings with *scale=TRUE* so the data is centered.

Support-Vector-Machine – SVM

The final model I used was the SVM. In my particular case, I focused on the linear version of the SVM. What a linear SVM does is try to find, using many different dimensions, the hyperplane that separates the data into the two classifications. However, since the data is most likely not linearly separable, we have the ability to say find a hyperplane that only accepts so much error. In the terms of this R model we call this the cost.

Since we can change the cost, this means this is a perfect model to use the k-fold cross validation with. This means we find the accuracy of classifying the training data for every cost, and the cost with the lowest error can be chosen. Then using the best model, we can test with out testing validation set and obtain our accuracy.

In order to implement SVM we will use the *svm* function and the *tune* function from the package *e1071*. The *svm* function will act as our SVM and the *tune* function is how we implement 10 fold cross validation. We will use cross validation with cost values of 0.00001,0.001,0.1,1, and10.

Results and Discussion

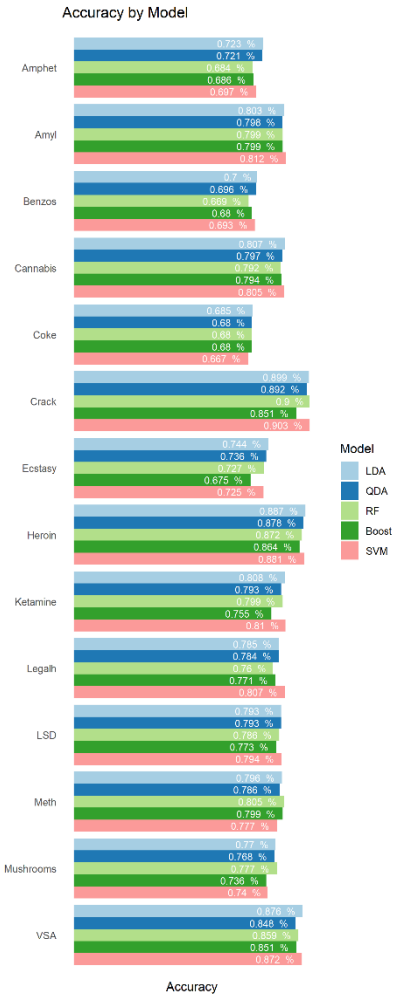

Figure 3. Accuracy results for Linear Discriminant Analysis, Quadratic Discriminant Analysis, Random Forest, Boosted Tree, and Support Vector Machine.

The results of the study can be seen in figure 3. Running each model on every drug we can see that some models perform better than others and some drugs are easier to model than others. Overall the models performed better than chance, which is good.

When it comes to drugs it appears that cocaine is the hardest drug to predict for. With every model hovering around 67/68% accuracy. Whereas a drug like Crack Cocaine is the easiest drug to predict. With the models landing near 90% accuracy. This begs the question of why this is. Is it perhaps something to do with data or model? Or is it some underlying connection with personality traits and drug usage. We cannot assume that personality traits effect drug usage, so it is possible that personality traits are not enough for Cocaine usage. Perhaps the price of Cocaine has something to do with this as well?

When it comes to models it appears that LDA did the best, followed by SVM. RF out performed only twice and QDA and Boosted tree did not outperform on any drug. Since LDA and SVM performed the best, it might seem that perhaps many of these drugs classifications are indeed linearly separable to some degree. Perhaps as well, the two drugs RF performed well on ,Methadone and Mushrooms, might be less linearly separable and may require a more flexible model to fit.

Overall the accuracy seems to be around 75%. Which seems to point to the idea that perhaps personality traits could be used as a way to predict if someone is likely to use a drug.

Conclusion

The goal of this study was to test whether or not using personality traits and basic demographics, if a model could predict whether a participant was a user of drug. The models used were LDA, QDA, RF, Boosted Tree, and SVM. From the results it seems like this could be likely. Given that each model did perform better than random chance this means that the personality traits does seem to contain useful information relating to drug use.

The model performed particularly well on the Crack Cocaine drug. This is interesting because it performed so badly on the Cocaine drug. Being that this is practically the same drug, but in different form it begs the question of why these performed so differently. This could possibly be an interesting question for future work, seeing what the models choose for the deciding factors between the two drugs. Perhaps, they could be different, or perhaps it could be due to something else such as the price difference between the two drugs.

Future work in this area could begin to tease apart these models and see if there is common scores between personality traits that relate to drug use. To see if there is a difference in traits between those that use drugs and those that don't and perhaps if there is one trait that is more heavily involved in the modeling.

Unfortunately, there was not enough time with project to explore all the interesting questions with this data set. However, perhaps in the future when I have learned more techniques I can return to this study and see if I can make improvements to the study and answer more of these interesting questions.

References

[1] A. Barry-Jester and A. Flowers. "How Baby Boomers Get High." *FiveThirtyEight*, https://fivethirtyeight.com/features/how-baby-boomers-get-high. Accessed 18 April 2021.

[2] Fehrman E, Egan V. Drug consumption, collected online March 2011 to March 2012, English-speaking countries. ICPSR36536-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-09-09. Deposited by Mirkes E. http://doi.org/10.3886/ICPSR36536.v1