# Machine Learning Analysis on Personality Type and Drug Consumption

Oscar Fawcett, Yesh Munagala, Emily Przykucki, Weston Murdock, and Ian Morton

## Abstract

In this paper, we explore the relationship between drug usage and different facets of personality and demographics by utilizing various machine learning classification techniques. Even though our implementation of k-Nearest Neighbors, Ordinal Logistic Regression, Naive Bayes, Neural Network, and Random Forest models all performed better than a random guess, our Ordinal Logistic models ranked highest in our analysis with a calculated test accuracy of ~ 0.69, then followed by Random Forest models with test accuracy of ~0.40. From this, we conclude that personality type has a moderate effect on drug usage.

# Motivation

In this analysis, we aim to create predictive models for drug consumption rates based on the personality of an individual, along with some demographic information. This is motivated by a desire to better understand the underlying reasons for why someone would use a given substance. If we can create an effective model in classifying drug usage, then our understanding behind substance use would increase through observing our most significant predictors as well as the types of models that work best. In essence, if a random forest model performs the best in classifying drug usage, that would yield a different interpretation on the reasons underlying drug usage than if a neural network performed the best.

# Data

To answer our research question, we explored the drug consumption dataset from the University of California Irvine Machine Learning Repository [1]. This dataset consists of 1885 observations and 31 variables, with 5 demographic predictors, 7 personality type predictors, and 18 reported drug consumption rate variables. The drug consumption variables are substances ranging from chocolate and caffeine to ketamine and heroin. Each drug has seven levels: CL0, CL1, CL2, CL3, CL4, CL5, and CL6, corresponding to never used, used over a decade ago, used in the last decade, used in the last year, used in the last month, used in the last week, and used in the last day, respectively. The demographic variables consist of age, gender, education, country, and ethnicity. Finally, the personality variables contain the Big Five Personality Type Indicators. Each of these five indicators corresponds to a score on a specific continuum. Openness relates to how curious and open to new experiences a person is; conscientiousness relates to impulsive v.s. goal-driven a person is; extraversion relates to how outgoing and adventurous a person is; agreeableness relates to how cooperative and trusting a person is; and neuroticism relates to an individual's emotional stability [2]. In addition to this, our dataset also contains the sensation seeking scores and impulsiveness scores for each respondent. These scores represent grades on how much the respondent seeks out sensory input and how impulsive they are, respectively.

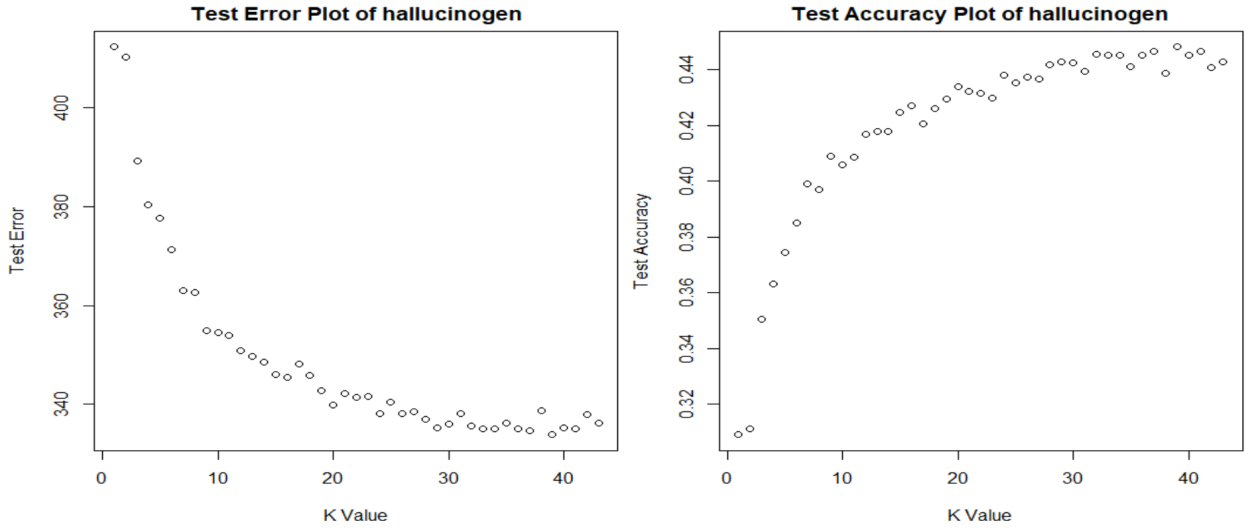
Due to a large number of response variables present in our dataset, we attempted to lessen computational complexity by creating 9 new variables for our models to classify: five variables in relation to the FDA's drug scheduling categorization, and three variables for drug effect type: stimulant, depressant, and hallucinogen, along with one additional variable for stimulants not including caffeine (we created this last variable because virtually all observations in our dataset use caffeine at high rates). These drug category variables were created by combining drug usage levels through 'or' statements, e.g., if a respondent answered 'used in the last day' to any of the drugs in the depressant category, then the depressant variable would hold the 'used in the last day' value. For the actual members in each of the nine above categories, the drug effect type variables are based on the body's response to taking a particular drug: stimulants speed up the function of the central nervous system, depressants slow down the function of the central nervous system, and hallucinogens affect the senses and change the way one sees, hears, tastes, smells or feels things [4]. The scheduling variables are based on the FDA's categorization of

drugs by their perceived medical usage and abuse potential. Schedule 1 are drugs with no currently accepted medical use and a high potential for abuse, schedule 2 are drugs with a high potential for abuse, with use potentially leading to severe psychological or physical dependence, schedule 3 are drugs with a moderate to low potential for physical and psychological dependence, and schedule 4 are drugs with a low potential for abuse and low risk of dependence [3]. We also included a schedule 0 category, which represents the drugs that are not scheduled, such as alcohol and caffeine.

Before we begin our analysis, it is important to note a few things from our exploratory data analysis: first, our personality data is all normally distributed. Second, our demographic data isn't very robust as it could be, i.e., variables such as country and ethnicity are very skewed towards the US and White, respectively, and age is recorded as a range as opposed to discrete numbers. And finally, our drug usage variables were very heavily skewed to one class or another, which represented an issue for us in creating accurate models.

## K-Nearest Neighbors

To begin our analysis, we implemented the k-nearest neighbors model to our data, using the drug scheduling and drug effect type variables as the responses. An important aspect of many machine learning techniques is the process of tuning hyperparameters. This was done for our k-nearest neighbors models by calculating the test accuracy for each k from 1 to the square root of our sample size, or 43. This process was completed using 5-fold cross-validation.




The above plot shows the calculated test error and accuracy for the hallucinogen response variable along with its corresponding k value. From this, we can see that we achieve an optimal k value at k = 39 with a test accuracy of approximately 0.44. This value was fairly standard across each of the response variables, with the average accuracy across each optimal model being approximately 0.52. However, as noted in the data section, our dataset is highly skewed, with certain classes of our response variables overrepresented. This issue could cause our models to give us skewed predictions, thus yielding an artificially high test accuracy. To combat this, we

randomly sampled from our dataset to guarantee that each class in our response variable was approximately evenly distributed and then trained our models on this new dataset. This was done by sampling a number of data points from each class equal to min(100, n), where n here represents the number of observations in a given class. We calculated test accuracy for this new dataset using 5-fold cross-validation, yielding the following results:

| Grouping (Unequal Training Split) | Average Accuracy |
|---|---|
| All Models: | 0.5232082 |
| Drug Scheduling Models: | 0.5102838 |
| Drug Classification Models: | 0.5361326 |
| Grouping (Equal Training Split) | Average Accuracy |
| All Models: | 0.2493896 |
| Drug Scheduling Models: | 0.2346087 |
| Drug Classification Models: | 0.2641706 |

Observing the above test accuracies, we can see that our models decrease efficacy by about 0.25 by using the equaling training split. Though a test accuracy of 0.25 isn't ideal, it is still an improvement over a random guess, which implies that our personality predictors do have an effect on drug consumption.

## Ordinal Logistic Regression

We proceeded with our analysis by performing ordinal logistic regression. We decided to use the following predictors: age, education, ethnicity, nscore, the Big Five Indicators and sensation seeking to build our nine models, five corresponding to the drug schedules (including the additional non-FDA schedule we created) and the latter four corresponding to the class of drugs. Mean accuracy across all nine models was ~0.577.

However, there exists class misrepresentation (arising from the highly skewed dataset), resulting in unusually lower accuracies. To mitigate this misrepresentation, we decided to use a similar test/train split as we did with the k-nearest neighbors algorithm. Indeed, we notice that average accuracy increased with the equal train/test split (unlike for some of the other methods we implemented) and conclude that the models are effective in predicting both the schedule and class of drug an individual is likely to use, as the accuracy is much better than a 'random guess'.

Mean test accuracy across all nine models are given in the below table:

| Grouping (Unequal Training Split) | Average Accuracy |
|---|---|
| All Models: | 0.5767757 |
| Drug Scheduling Models: | 0.5893899 |
| Drug Classification Models: | 0.4488064 |
| Grouping (Equal Training Split) | Average Accuracy |
| All Models: | 0.6894098 |
| Drug Scheduling Models: | 0.713097 |
| Drug Classification Models: | 0.6598008 |

## Naive Bayes

In our next step of the analysis, we created Naive Bayes models to classify which personality traits and demographic components corresponded to certain drug usage classes. In order to implement this model, we had to assume that the predicting features are independent of one another and that they all have an equal influence on the outcome, as these are conditions for Naive Bayes models.

Because we deemed it likely that different personality traits will correspond to propensities to use different types of drugs, we decided to make four separate Naive Bayes models - one each to predict usage of stimulants, hallucinogens, depressants, and stimulants excluding caffeine. We used all demographic variables and all personality variables in the creation of these models. Lastly, we made training and test data with an 80%/20% split.

The first time that we created these models, we did so with the ‘raw data’. That is, we used the data in its entirety, despite it having some over/underrepresented usage classes. Much like was the case with the k-nearest neighbors method, these models produced artificially high accuracies, since the model could use the polarization of the data as a crutch. In other words, because of the nature of certain drugs to be highly addictive, most observations were either clustered in CL0/1 or CL5/6, which rewarded the model for disproportionately predicting someone to fall in these categories. To combat this error, we re-created the models; this time with equal representation from each drug class. While our accuracies did decrease, we believe this adjustment improved the models through the elimination of error and/or bias. The table below displays our model accuracies, both before and after our adjustment.

| | | ACCURACIES | | |
|---|---|---|---|---|
| | | "Raw" data | Data w/ 100 from each usage class | Change |
| MODELS | **Stimulants** | 0.813 | 0.347 | **-0.466** |
| | **Depressants** | 0.5087 | 0.336 | **-0.1727** |
| | **Hallucinogens** | 0.5112 | 0.3557 | **-0.1555** |
| | **Stimulants** (no caff) | 0.419 | 0.1308 | **-0.2882** |

The steepest decrease after the class distribution adjustment was 'Stimulants' (going from 81.3% accuracy to 34.7%). This can be attributed to caffeine - a drug that is frequently consumed by most people, causing a severe skew in the data. Additionally, because caffeine consumption is so common and widely accepted, there are no specific personality traits or demographic qualities that can be reliably associated with it. With seven different usage classes, randomly guessing would yield ~14% accuracy. For stimulants, depressants, and hallucinogens (all after adjustments), our accuracy levels fell decently high above this benchmark (around 35% for all). However, the accuracy for the stimulant group excluding caffeine did not necessarily exceed a random guess level, meaning the Naive Bayes model for this category of drug cannot reliably predict usage class.

## Neural Network

Next, we applied neural network models to our 9 created response variables, using the Big Five Personality Type Indicators as predictors. We trained neural network models with one, two, and three hidden layers each with a ReLU activation. After attempting unit values of 8, 16, 32, and 64, we saw the best neural network model had one hidden layer and 16 units. In addition, we used the RMSprop optimizer with a categorical cross-entropy loss. These models yielded an average accuracy of approximately 0.52 across all nine models. Furthermore, just like have done in previous sections, we subsetted the data into evenly distributed classes to account for class overrepresentation. This resulted the following accuracy table:

| Grouping (Unequal Training Split) | Average Accuracy |
|---|---|
| All Models: | 0.5105932 |
| Drug Scheduling Models: | 0.5042373 |
| Drug Classification Models: | 0.5169492 |
| **Grouping (Equal Training Split)** | **Average Accuracy** |
| All Models: | 0.2458204 |

| | |
|---|---|
| Drug Scheduling Models: | 0.2316172 |
| Drug Classification Models: | 0.2600236 |

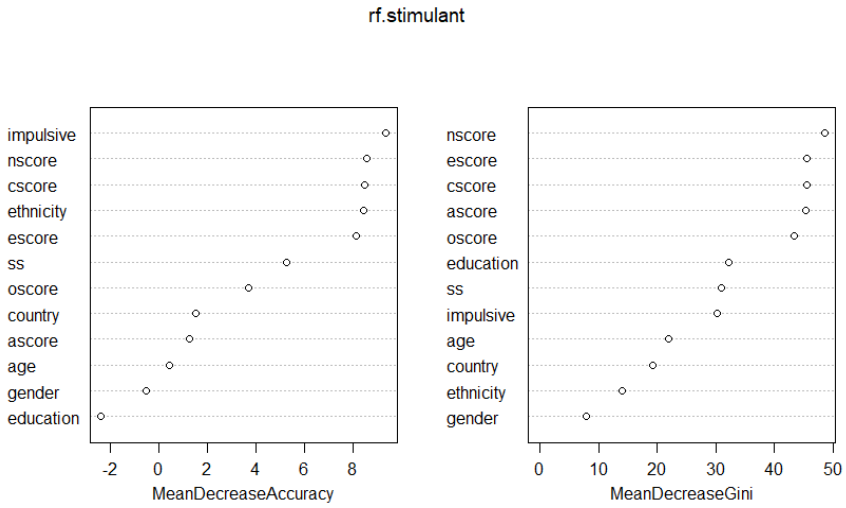
As we can see from the above table, our neural network doesn't perform especially well at classifying our dataset compared to that of our previous models. That being said, it is still an improvement over a random guess.

## Random Forests

Next, using decision trees, we created a bagging and random forest model for every individual drug, every drug schedule, and every drug classification in our response variables, using all of the demographic and personality variables as predictors for the response variable of the last time a drug was used. For the bagging models, all 12 of these predictor variables were used to generate each tree, with 500 trees generated to arrive at the final model, while the random forest models all used a random selection of four of the predictor variables to generate each tree, also using 500 trees to arrive at the final model. The data was divided into a training set of 1,202 observations and a testing set of 683 observations out of the total 1,885 observations for the purposes of creating and testing each of these models, with the training and testing set remaining the same across every one of the models for consistency. Additionally, as was done for the previous models, after creating an initial bagging and random forest model for each response variable, we then created an additional version of each model for the drug scheduling and classification response variables that were trained on a data set that included an equal number of observations from each of the response classes (CL0 through CL6). Since there were a total of 28 response variables that models were initially created for and nine response variables that equally sampled models were created for, with each having a bagging and random forest model created, this resulted in a total of 74 decision tree models being created. A summary of the results of these models is shown in the table below, along with the importance plot of the predictors for the random forest model generated for the stimulant classification.

| | |
|---|---|
| Best Model: | Crack (RF) (Accuracy: 0.85) |
| Worst Model: | Cannabis (Bagging) (Accuracy: 0.37) |
| Grouping (Unequal Training Split) | Average Accuracy (Bagging/RF) |
| All Models: | 0.591 / 0.600 |
| Drug Scheduling Models: | 0.558 / 0.572 |
| Drug Classification Models: | 0.537 / 0.550 |
| Grouping (Equal Training Split) | Average Accuracy (Bagging / RF) |

| | |
|---|---|
| All Models: | 0.420 / 0.432 |
| Drug Scheduling Models: | 0.468 / 0.483 |
| Drug Classification Models: | 0.360 / 0.369 |




From these results, we may conclude that random forest models generally performed better than bagging models, that both types of models performed generally worse on groupings of drugs such as by classification or schedule than on individual drugs (in most cases), and that response variables with a more even distribution of responses were more difficult to create accurate models for (as indicated by the poor performance of the cannabis models). This last point may in some cases also help to explain why the models performed worse on the groupings of drugs since these variables often had a wider spread in the responses than the individual drugs within the grouping on their own. Furthermore, the importance plot above shows that while demographic variables can be useful in predicting a person's drug usage habits, the personality variables were frequently the most important predictors, even for groupings of drugs like the stimulant group, supporting our overall goal of using these personality measurements to predict drug usage habits.

Finally, future versions of these models could potentially improve upon the originals (or at least provide more insight) by mapping all of the response classes to a set of numbers, training the models as usual, and then using the numbers to calculate the loss of the models in a separate manner from the accuracy, such as mean squared error, in order to get a better idea of not just whether or not the model is sorting people into the right category for drug usage, but also how far away the model is guessing when it guesses incorrectly.

# Conclusion

The primary purpose of our analysis was to determine if there is a detectable correlation between certain facets of personality or demographic traits and the frequency of drug use. To

answer this, we trained and tested our data on various models. Most models yielded classification accuracy modestly above a 'random guess' level. However, our ordinal logistic regression and random forests models yielded fairly high results, with test accuracies of approximately 69% and 40% respectively. Though these accuracy levels demonstrate that our predictor variables can be used to anticipate drug usage, we believe that there is likely a myriad of other factors that contribute to drug usage, some of which may not even be measurable. We also attempted Linear Discriminant Analysis and Quadratic Discriminant Analysis but they produced accuracy levels far lower than that of a random guess. As such, these values are not included in this report. Below table summarizes mean test accuracy across different techniques:

| KNN | Logistic Regression | Random Forests | N-N | Naive Bayes |
|---|---|---|---|---|
| ~0.22 | ~0.69 | ~0.43 | ~0.25 | ~0.29 |

Additionally, the above techniques revealed that, among our variables, personality traits were generally more useful in predicting drug usage. However, if we were to repeat this analysis with access to more information, we would like to have more in-depth demographic variables (for example: income, place of birth, family history, etc.) to include in our models. The relationship between demographic traits and personality traits as they relate to drug usage is nuanced - which ones cause drug use and which ones are merely correlated with drug use? Or, how does the presence of some of these variables potentially influence the presence of others? Drug consumption is an inherently complex matter, and more information on multicollinearity between variables would provide a lot of insight into what is a true predictor of drug consumption.

# Works Cited

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

[2] https://www.simplypsychology.org/big-five-personality.html

[3] https://www.dea.gov/drug-information/drug-scheduling

[4] https://www.health.gov.au/health-topics/drugs/about-drugs/types-of-drugs