CAN PERSONALITY FEATURES PREDICT SUBSTANCE ABUSE ?

César René Pabon Bernál

Capstone Project (Springboard)

March 29th, 2018

CAN PERSONALITY FEATURES PREDICT SUBSTANCE ABUSE ?

César René Pabon Bernál
Springboard

March 29th, 2018
Capstone Project

CAN PERSONALITY FEATURES PREDICT SUBSTANCE ABUSE ?

1. INTRODUCTION

For most New York City middle school public school students, summer school hours tend to be boring, uncomfortably hot and never ending. Seconds feel like minutes, hours feel like days—such was my only memory of sixth grade. The last school day was meant to be exciting and joyous, however, it was instead greeted by a busy calendar full of seminars covering subjects like drug abuse and the now infamous D.A.R.E. program (Drug Abuse Resistance Education). Surprisingly, I paid attention, and the widely state-adopted substance abuse prevention initiative (operational from 2001-2007), DARE and its purpose, resonated with me not only that day, but has been ever since. I can state, that DARE positively delivered valuable information to my awareness network, and therefore can reason that the program was effective in contributing to my overall drug use defense system. Unfortunately, for most of my classmates who were present during these seminars, their reaction was different. At the time, there were too many questions, and not enough answers to reasons why drug use prevention ideas stuck with me but didn't for others: A precursor for substance abuse, was the desire for drug experimentation stronger in one person versus the other? If so, why? Is it because our personalities are different? Was it due to gender and/or other demographics?[1]

As of 2017, studies have found that substance abuse prevention programs, such as DARE, were largely ineffective due to their dependence on "scare tactics." This created an opposite target effect, leaving children highest at risk, vulnerable when dealing with possible life changing scenarios.[2] Continued research has focused in understanding the precursors that can lead to drug abuse with intentions in discovering a possible "root" to the problem—Something that is incredibly vital to the progression of a healthy society.[3]

[1] Kumar, Revathy et al. "Alcohol, Tobacco, and Other Drug Use Prevention Programs in U.S. Schools: A Descriptive Summary." *Prevention science : the official journal of the Society for Prevention Research* 14.6 (2013): 581–592. *PMC*. Web. 19 Mar. 2018.

"Monitoring the Future National Survey Results on Drug Use, 1975-2017: Overview, Key Findings on Adolescent Drug Use." 2018, doi: 10.3998/2027.42/142406.

[3] Szalavitz, Maia. "The 4 Traits That Put Kids at Risk for Addiction." *The New York Times*, The New York Times, 29 Sept. 2016, www.nytimes.com/2016/10/04/well/family/the-4-traits-that-put-kids-at-risk-for-addiction.html.

2. CLIENT AND PROBLEM

According to the National Institute of Drug Abuse, abuse of psychoactive drugs is a devastating global epidemic consisting of complex factors; environmental, economic, and social being the most important. Without proper treatment, these can have devastating direct (increase in risk of poor health) and indirect consequences affecting families, friends, and entire life networks.[4,5] On October 26th of 2017, the president of the United States of America "directed the Department of Health and Human Services to declare the opioid crisis a public health emergency."[6] Government agencies, educational institutions, and private companies continue to dedicate insurmountable resources to tackle this problem—it costs the United States of America over $600 billion annually.[7]

Over the last decade, great progress has been made towards understanding the root of substance abuse. A recent scientific article was published from Cornell University by a group of respected statisticians and psychologists, and found that there is a correlation between psychoactive drug consumption and the "Big Five Personality Traits," also known as the fingerprint to our behavioral patterns. Our objective in this review is to analyze the anonymous online database survey utilized in the above publication and focus on two points:

1) The potential effect between the "Big Five Personality Traits," demographic data and drug consumption for different psychoactive drugs.

2) Predict the probability binary correlation of drug consumption for each respondent based on

a. personality attributes vs all psychoactive drugs
b. personality attributes vs the opioids cluster (heroin, methadone, and legal highs)[8]

[4] McGinnis JM, Foege WH. Actual causes of death in the United States. Journal of the American Medical Association. 1993; 270(18):2207–2212.

[5] Sutina AR, Evans MK, Zonderman AB. Personality traits and illicit substances: the moderation role of poverty. Drug and Alcohol Dependence. 2013; 131:247–251.

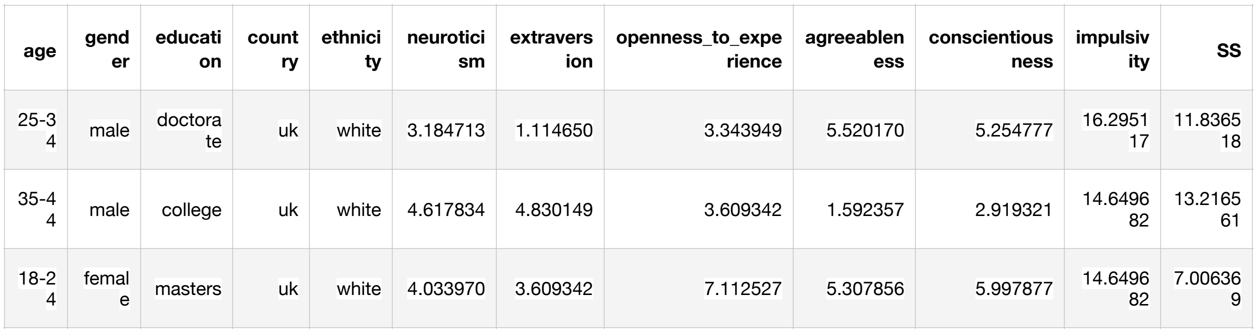
[6] Davis, Julie Hirschfeld. "Trump Declares Opioid Crisis a 'Health Emergency' but Requests No Funds." *The New York Times*, The New York Times, 26 Oct. 2017, www.nytimes.com/2017/10/26/us/politics/trump-opioid-crisis.html?hp.

[7] NIDA. "Principles of Drug Addiction Treatment: A Research-Based Guide (Third Edition)." *National Institute on Drug Abuse*, 17 Jan. 2018, https://www.drugabuse.gov/publications/principles-drug-addiction-treatment-research-based-guide-third-edition. Accessed 19 Mar. 2018.

[8] Fehrman, et al. "The Five Factor Model of Personality and Evaluation of Drug Consumption Risk." *[1506.06297] The Five Factor Model of Personality and Evaluation of Drug Consumption Risk*, 15 Jan. 2017, arxiv.org/abs/1506.06297.

3. DATA

Data was downloaded from the UCI Machine Learning Repository portal and it's composed of answers to an anonymous online drug use questionnaire (administered by the first author in the referenced publication, Dr. Elaine Fehrman). The raw data itself was in a CSV format, mislabeled and disorganized. An appropriate dictionary was created and uploaded to replace each feature/trait in order to reflect their true value. It contains 1884 respondents (941 females and 943 males), 12 personality attributes about each participant (7 are behavioral personality traits) and information on consumption for 18 central nervous system psychoactive drugs. The dimension of the final clean dataset is 1884 X 32.



age	gender	education	country	ethnicity	neuroticism	extraversion	openness_to_experience	agreeableness	conscientiousness	impulsivity	SS
25-34	male	doctorate	uk	white	3.184713	1.114650	3.343949	5.520170	5.254777	16.295117	11.836518
35-44	male	college	uk	white	4.617834	4.830149	3.609342	1.592357	2.919321	14.649682	13.216561
18-24	female	masters	uk	white	4.033970	3.609342	7.112527	5.307856	5.997877	14.649682	7.006369

A) 12 personality attributes

1. Age
2. Gender
3. Education (level of education of participant)
4. Country
5. Ethnicity
6. Openness_to_experience
7. Conscientiousness
8. Extraversion
9. Agreeableness
10. Neuroticism
11. Impulsiveness
12. Sensation Seeking (S.S)

O.C.E.A.N stands for the "Big Five Personality Traits"

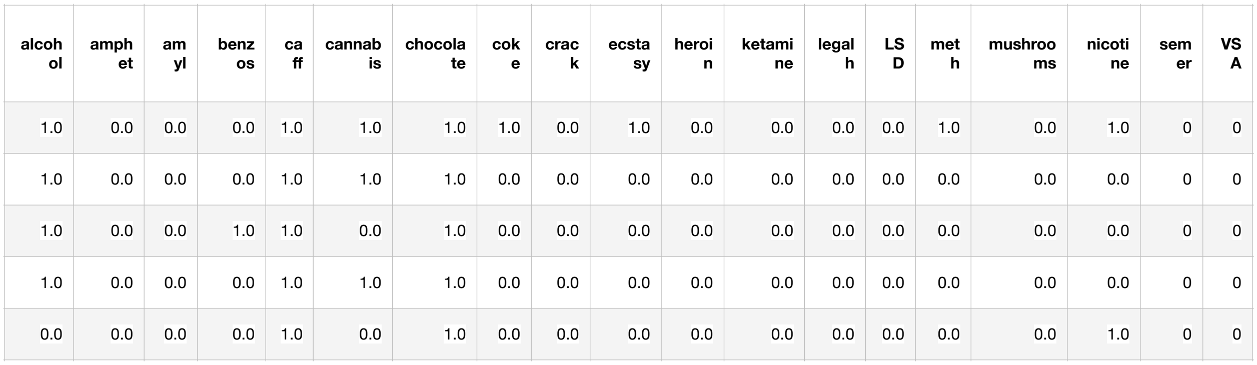
It is important to state, that behavioral personality trait values (No. 6-12) are weighted answers which were then transformed into a percentage.

B) 18 central nervous system psychoactive drugs

1. Alcohol
2. Amphetamines Amyl nitrite
3. Amyl nitrite
4. Benzodiazepines
5. Caffeine
6. Chocolate
7. Cocaine powder (coke)
8. Cocaine solid (crack)
9. Ecstasy
10. Heroin
11. Ketamine
12. Legal highs
13. LSD
14. Methadone
15. Magic mushrooms
16. Nicotine
17. Semeron (Fictitious and introduced to identify over-claimers)
18. VSA (Violent solvent abuse)

Opioids Cluster: No. 10 Heroin, No. 12 Legal highs,
No. 14 Methadone

A dummy variable system was created for a respondents drug consumption use. A zero "0" was assigned to those who have never taken the drug in question or used it over a decade (Non-User). A one "1" was assigned to those who have used the drug in question during the last day/last week/last month/last year /last decade (User).



alcohol	amphet	amyl	benzos	caff	cannabis	chocolate	coke	crack	ecstasy	heroin	ketamine	legalh	LSD	meth	mushrooms	nicotine	semer	VSA
1.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0	0
1.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0
1.0	0.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0
1.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0	0
0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0	0

There are 3 major type of drugs; depressants (alcohol, amyl nitrite, benzodiazepines, heroin, methadone, legal highs, VSA), stimulants (nicotine, coke, crack, caffeine, chocolate, amphetamines), and hallucinogens (cannabis, ecstasy, ketamine, LSD, magic mushrooms).

4. EXPLORATORY/STATISTICAL ANALYSIS

According to our client problem, we are to analyze the anonymous online database survey and focus on two parts:

1. The potential effect between the "Big Five Personality Traits," demographic data and drug consumption for different psychoactive drugs.

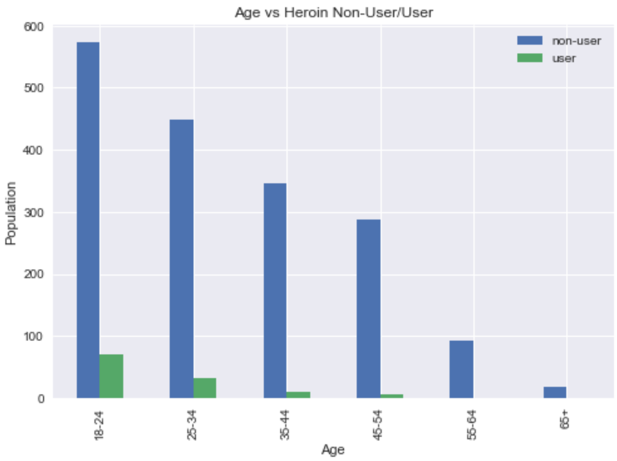
2. Predict the probability binary correlation of drug consumption for each respondent based on

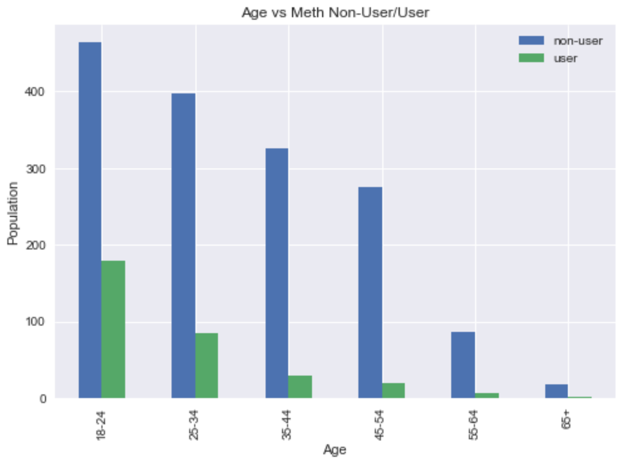
   a. Personality attributes vs All psychoactive drugs
   b. Personality attributes vs Opioids cluster (heroin, methadone, and legal highs)

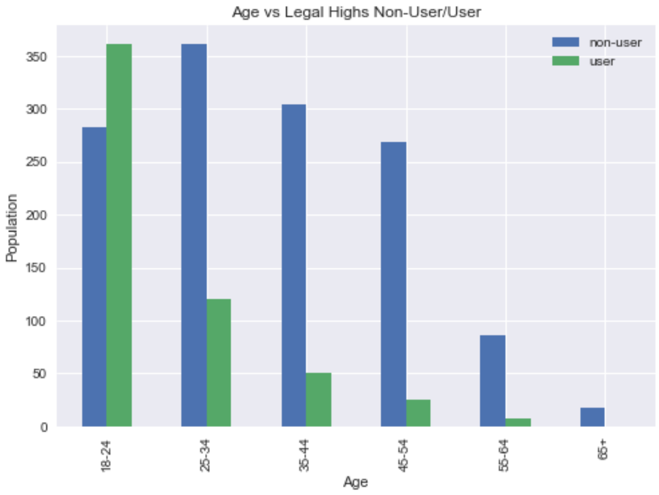
We will formulate appropriate exploratory graphs and hypothesis testing based on the conclusions found in the reference publication and our own study. Each of the 12 personality features/traits (categorical and non categorical) play an important role when creating a unique and complex profile for each respondent. However, some have more weighted value than others towards outlining the root of the problem. We will examine the following:

i. Age vs. Opioids Cluster
ii. Highest Education Achieved vs vs. Opioids Cluster
iii. Country of Origin vs. Opioids Cluster
iv. Ethnicity vs. Opioids Cluster
v. Gender vs. vs. Opioids Cluster
vi. Neuroticism & Openness_to_experience trait vs Heroin Non-User/User

i. Age vs. Opioids Cluster



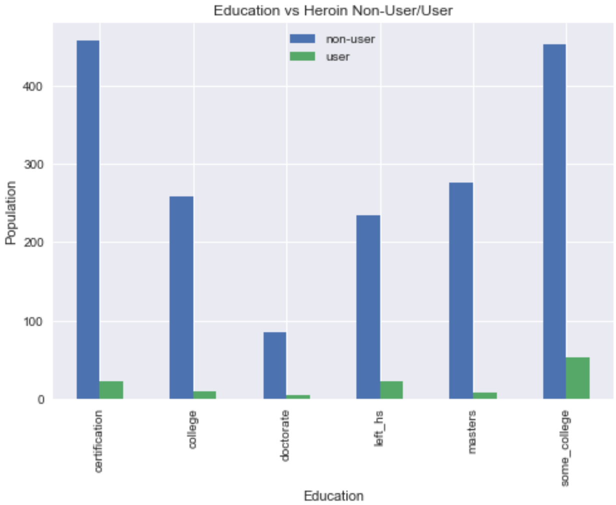


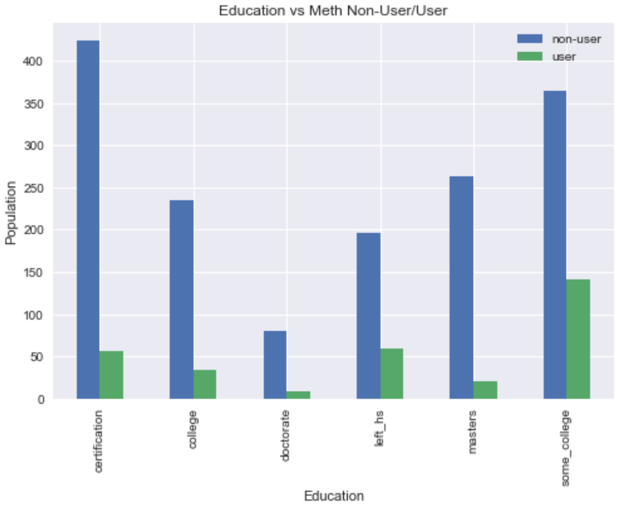


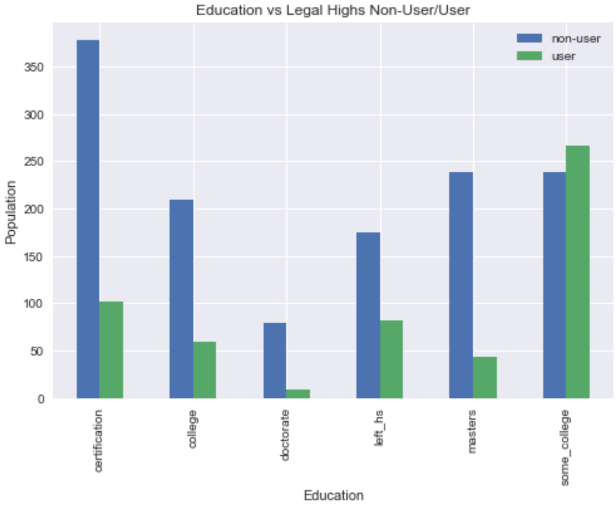
Age	% of people using legal highs
All	29.94
18-24	19.16
25-34	6.37

From the Age vs. Opioids section, we are able to conclude two things. Firstly, most drug users range from ages 18-24. And secondly, abuse of legal highs (drugs that contain various chemical ingredients that produce similar effects to illegal and legal stimulants) is cumulatively higher than the use of heroin and methadone.

ii. Highest Education Achieved vs. Opioids Cluster



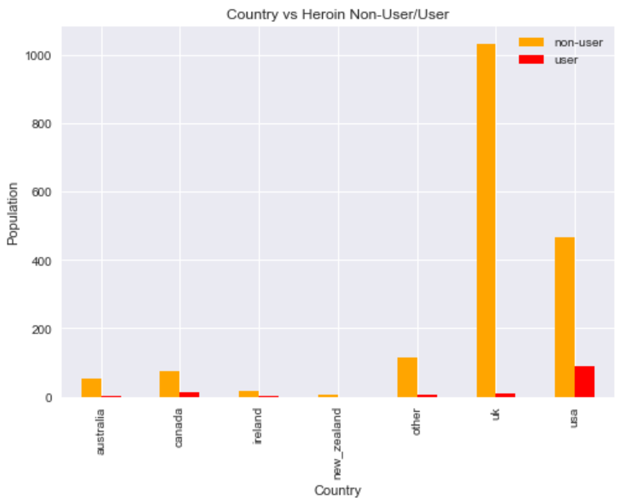


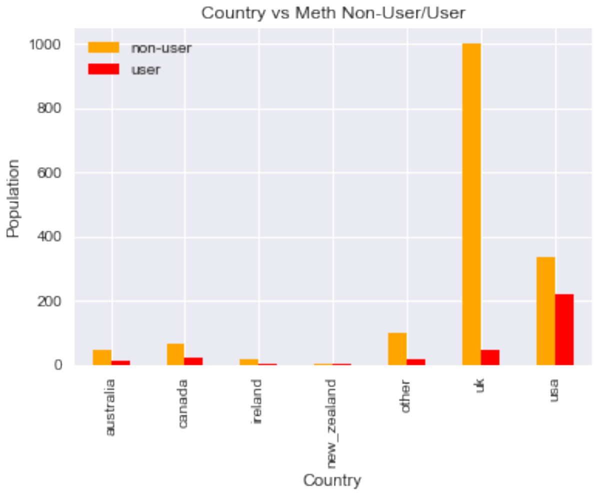


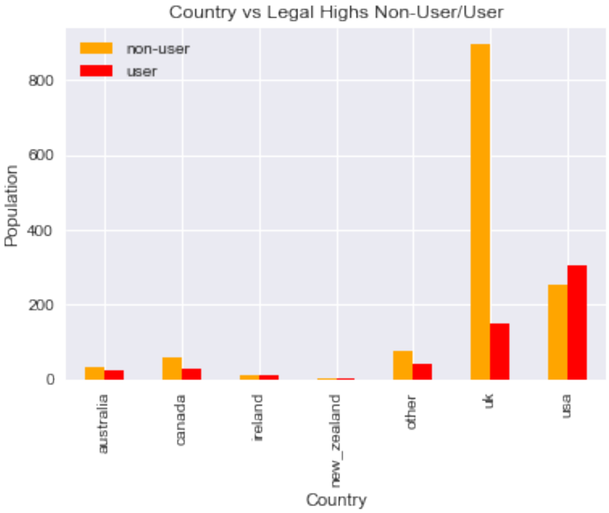
Education	% of people using legal highs
All	29.94
some_college	14.17
left_HS	4.35

From the Education vs. Opioids section, we are able to conclude two things. Firstly, most drug users have at least some college education (at least of two years completed in a four year college program). And secondly, abuse of legal highs is cumulatively higher than the use of heroin and methadone.

iii. Country of Origin vs. Opioids Cluster







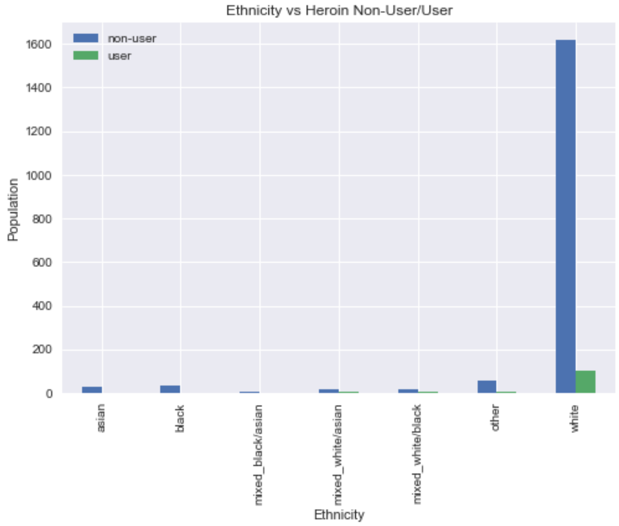
Country	% of people using legal highs
All	29.94
USA	16.14
UK	7.91

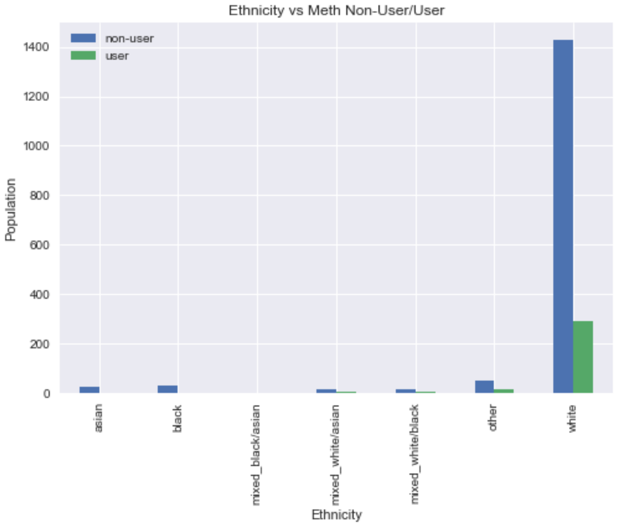
From the Country vs. Opioids section, we are able to conclude two things. Firstly, most drug users are from USA or UK. And secondly, abuse of legal highs and methadone is about the same level. An incredibly visually alarming statement as meth appears to be increasing in use in the US. This confirms a recent National Drug Threat Assessment summary conducted by the United States Drug Enforcement Administration.[9,10]

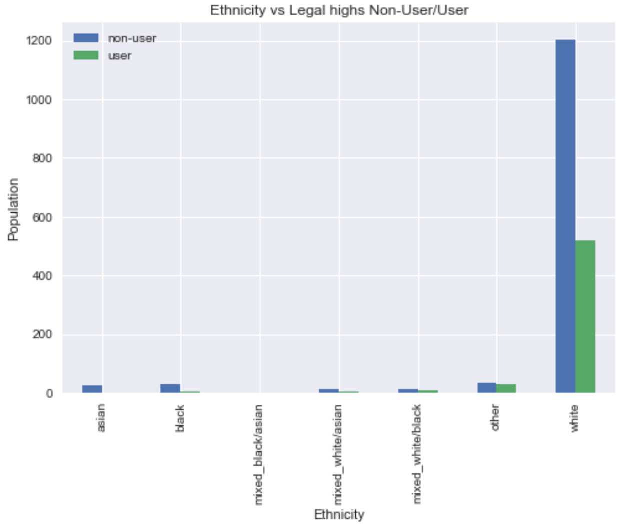
[9] “2016 National Drug Threat Assessment Summary.” *Homeland Security Digital Library*, United States. Drug Enforcement Administration, 1 Nov. 2016, www.hsdl.org/?abstract&did=797265.

[10] “Meth Use Surges in Western, Southern and Midwestern States.” *Governing Magazine: State and Local Government News for America's Leaders*, www.governing.com/topics/public-justice-safety/sl-meth-usage-drugs-states-illegal.html.

iv. Ethnicity vs. Opioids Cluster





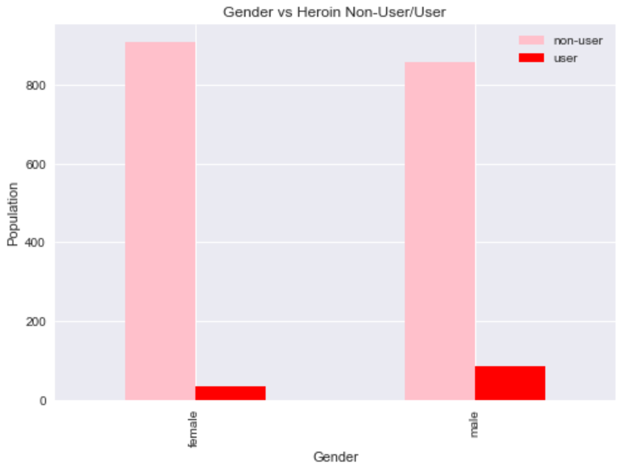


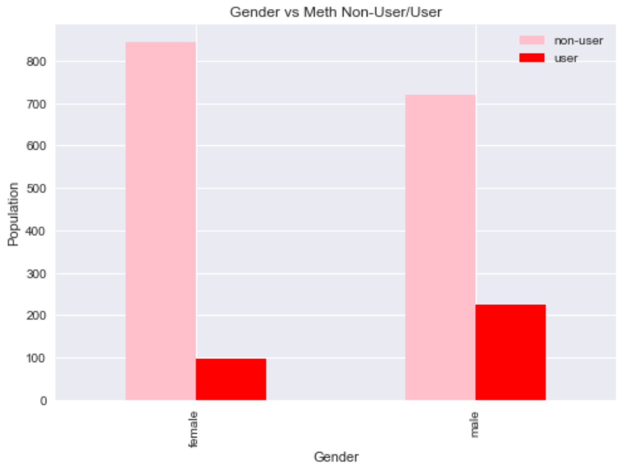
Ethnicity	% of people using legal highs
All	29.94
White	27.49
Other	1.59

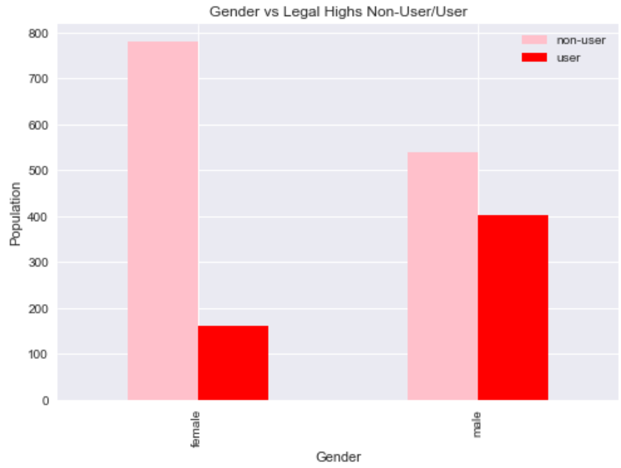
From the Ethnicity vs. Opioids section, we are able to conclude that most drug users are of white ethnicity by a substantial margin. This confirms a 2016 study conducted by The Henry J. Kaiser Family Foundation in which 33,450 overdose deaths, out of 42,429, where reported as White-Non Hispanic.[11]

[11] "Opioid Overdose Deaths by Race/Ethnicity." *The Henry J. Kaiser Family Foundation*, https://www.kff.org/other/state-indicator/opioid-overdose-deaths-by-raceethnicity/?currentTimeframe=0&sortModel=%7B%22colId%22%3A%22Location%22%2C%22sort%22%3A%22asc%22%7D.

v. Gender vs. vs. Opioids Cluster



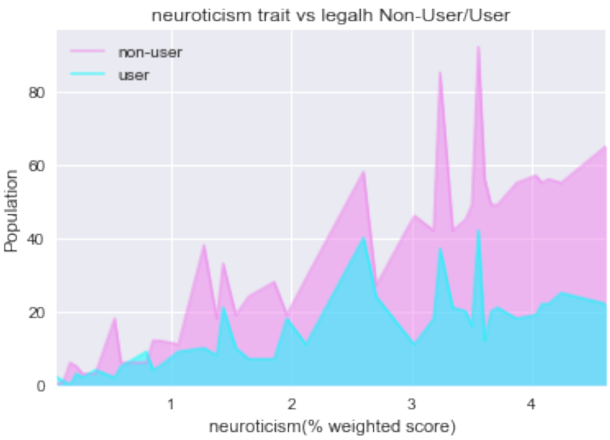


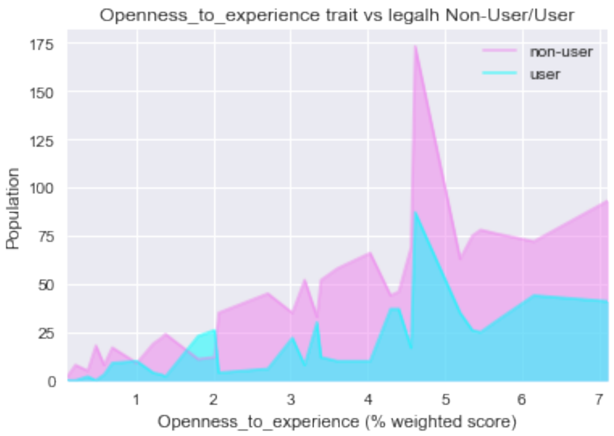


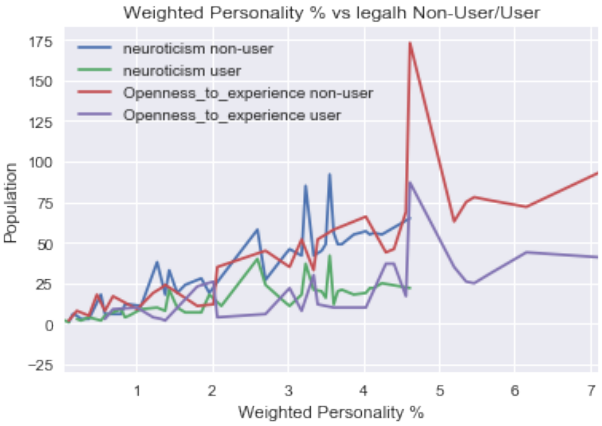
Gender	% of people using legal highs
All	29.94
Male	21.39
Female	8.55

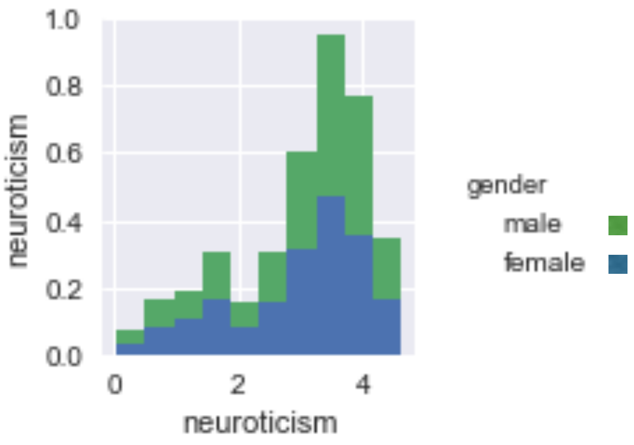
From the Gender vs. Opioids section, we conclude that most drug users are male, by about a two to one ratio. This confirms the same Henry J. Kaiser Family Foundation 2016 study from the Ethnicity vs. Opioid section in which 28,498 overdose deaths were male and 13,751 were female.

vi. Neuroticism & Openness_to_experience trait vs Heroin Non-User/User









Our exploratory analysis on behavioral traits confirm the referenced paper. Neuroticism and openness to experience score the highest weighted personalty values when determining whether someones is a drug use or not—especially legal highs which is the most vulnerable of the depressant opioid cluster.

vii. SUMMARY

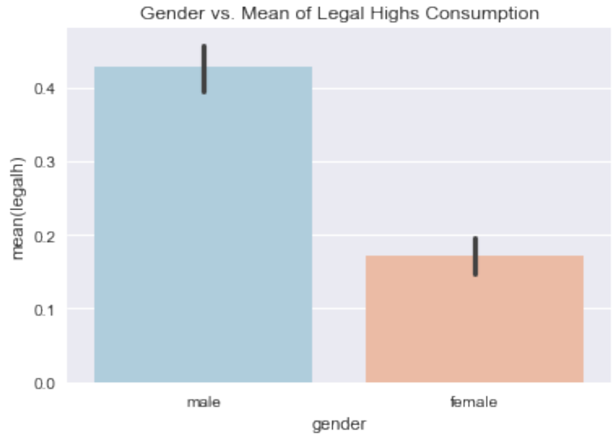
It appears that the most vulnerable personality profile for substance abuse is a person that is an American white male, between the ages of 18-24, with some college education with high levels of neuroticism and openness to experience. Based on the visual exploratory analysis, we can now interpret the results to create hypothesis testing.

INFERENTIAL STATISTICAL ANALYSIS

We will evaluate gender versus legal highs and state that male and female populations are independent with an assumption that normality was violated. A 2-sample t hypothesis test is appropriate in determining which gender is more susceptible of using legal highs. We will:

1. Determine whether the means of male and female differ.
2. Calculate a range of values that is likely that include the difference between the population

i. HYPOTHESIS TESTNG



Gender	Total Population	Population using legal h	% using legal h
All	1884	564	——
Male	943	403	71.44
Female	941	161	28.55

As we can see, the percent difference a male will use legal highs versus a female is 42.90% and validates the Gender vs Mean of legal high consumption bar-plot. By looking at the population of the samples we can quantify precision as an estimate of the effect size through a hypothesis test.

Null hypothesis

H0: $\mu1- \mu2 = \delta0$

- There is *no* difference between the gender/population means/proportions for both groups on deciding whether they use legal highs.

Alternative hypothesis

H1: $\mu 1- \mu 2 \neq \delta 0$.

- There *is* a difference between the gender/population means/proportions for both groups on deciding whether they use legal highs (does not equal zero).

Results

After performing a t-test, our p-value score was 3.7987E-35. The small P-value provides a statistical significance when evaluating the difference between the population means for both groups and their respective decision on whether to use legal highs or not. Therefore, we reject the null hypothesis in favor of the alternate hypothesis.

ii. STATISTICAL ANALYSIS SUMMARY

Gender	Mean %	Margin of Error %	Confidence Interval (low %)	Confidence Interval (high %)
Male	71.45	2.04	69.41	73.49
Female	28.55	2.04	26.51	30.59

In a group composed of 943 males and 941 females, we report with a 95% confidence, that:

1. Males use legal highs 69.4 - 72.5% of the time; or 71.45 ± 2.04% of the time.
2. Females use legal highs 26.5 - 30.6% of the time; or 28.55 ± 2.04% of the time.

Our C.I. findings are in proportion with the results in the published paper. With a 95% confidence interval it was reported, that for all genders, use of legal highs ranged 47.50% and 50.32%. We use this conclusion as a lead way into the evaluation of a more complex system. By employing classification analysis, we can predict the probability binary correlation of drug consumption for each individual based on all personality features/traits.

5. MACHINE LEARNING AND PREDICTIVE MODELING

We evaluate the second part of the client problem and ask: can we predict the probability of drug consumption for everyone based on

a. Personality attributes vs all psychoactive drugs
b. Personality attributes vs the opioids cluster (heroin, methadone, and legal highs)

The first step when evaluating complex datasets with a population size greater than 30 is to determine if we are able to first, establish a target problem and second, identify its type. In the exploratory statistical inference section, we concluded that a big part of the dataset is categorical, qualitative and capable of accepting binary questions/responses. In order to predict the probability of drug consumption for any respondent we have decided to use logistic regression (a supervised machine learning method). Using logistic regression, we are able to predict

a. Response of Y as binary (someone uses a drug when yes=1 or no=0)
b. Probability of Y

We restate that the dataset contains 1884 total respondents (941 females and 943 males), 12 personality attributes about each participant (7 are behavioral personality traits) and information on consumption for 18 central nervous system psychoactive drugs. Utilization of a "dummies" system (X/Y matrix) was created in order to convert the qualitative features (age, gender, education, country, and ethnicity) into dummy/indicator variables—useful for applying some machine learning algorithms.

A. LOGISTIC REGRESSION MODEL

As we build our binary classification model (using python), we do so, with the requirement that it not only fits our current dataset well, but also, new unseen data. The general layout for this type of model is to first split the entire dataset in two parts: a training set and a test set. Second, the training set is then fitted or trained. And third, the information gathered from the fitted trained set is tested for accuracy iterated over the testing set. We will evaluate these steps using three separate techniques:

1. K-fold Cross Validation w/regularization parameter C
2. Grid Search Cross Validation w/regularization parameter C
3. Train/Test Data Splitting Cross Validation w/out regularization parameter C

All three methods use cross-validation which robustly estimates the test-set performance of the model. K-fold cross-validation splits the data into *k*-bins, Grid-search selects the best of a family of models, while test/train simply splits the data in two. Both K-fold and Grid Search cross-validation share a very important parameter, the regularization parameter C. Some binary classification models have hyperparameters that can be tuned for better performance. In our Logistic Regression model, the most important parameter to tune is the regularization parameter C. This parameter controls for unlikely high regression coefficients and can be used as a method for feature selection when data is sporadically dispersed. In order to determine which model works best to predict the probability binary correlation of drug consumption for each individual, we followed these steps:

Tuning model with K-fold cross validation

- Use the cv_score function to perform K-fold cross-validation.
- Use the scoring function to test each fold.
- Through the now optimized cv_function, we find the best model C parameters based **only** on the training set.
- We create a new logistic regression model with the newly calculated best regularization parameter C and train, only ,again, on the training data.
- Finally, we train logistic regression (using the best regularization parameter C and the predictive function) on **test** data
- We use the accuracy_score function to test on the models accuracy for predicting y values.

The system model creates K train/test bins and iterates the best averaged results as loops over model parameters. K-fold tends to be a better estimate for an out-of-shape sample performance than test/train splitting, however, it tends to run K times slower.

Tuning model with Scikit-learn's Grid Search Cross Validation

- Use the grid search function to apply logistic regression on the parameters using cross-validation
- Fit the training set, **not** the test data, and find the best
    1. Estimator C and its respective parameters
    2. Best score for the test data and confirmation of chosen parameters
        i. Mean scores of grid scores
        ii.. Standard Deviation
        iii. parameters
- Use the best regularization estimator C found in the second step to fit the **test** data
- Use the accuracy_score function to test on the models accuracy for predicting y values.

The Grid Search cross-validation instance implements the estimator API when "fitting" onto the dataset using all possible combinations of parameter values (parametrization). This is then evaluated and the best combination is retained. Grid Search is typically the better estimate for of out-of-shape sample performance between k-fold CV and test/train splitting.

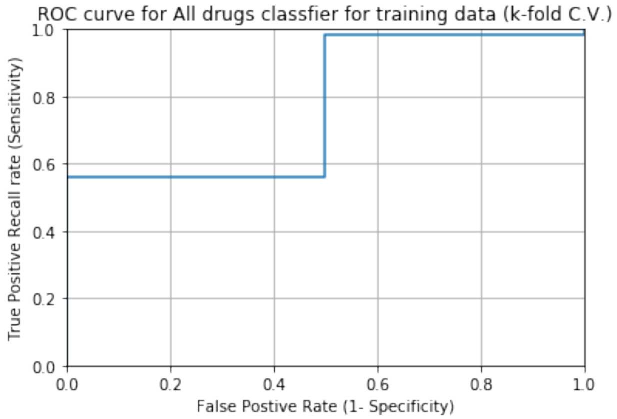
<u>Train/Test Splitting Cross Validation</u>

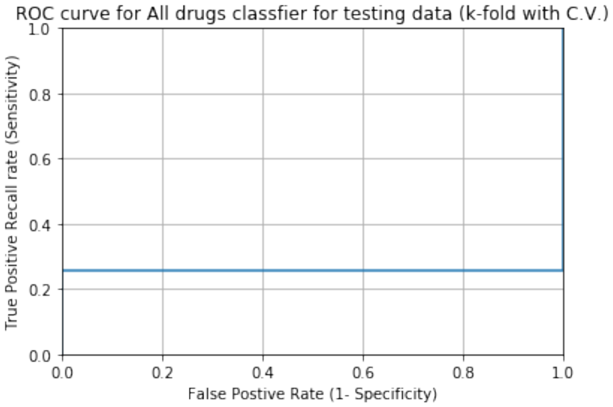
- Use the train_test_split function to split the data in two sets
- Fit the data using the L.R. function
- We use the accuracy_score function to test on the models accuracy for predicting y values.

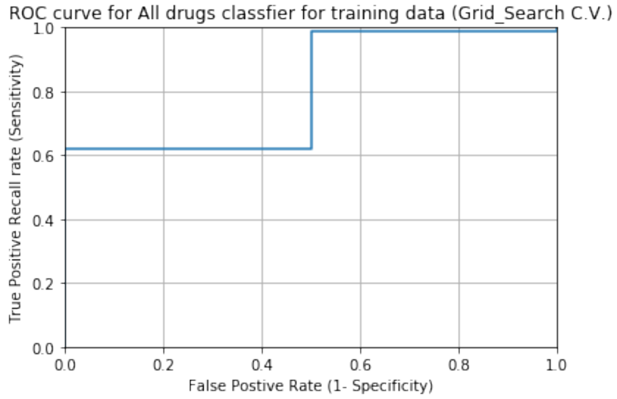
Train/test splitting allows the model to be trained and tested, independently. It is fast, simple, flexible and creates a good estimate on out-of-sample performance. However, it generates a **high** variance estimate.

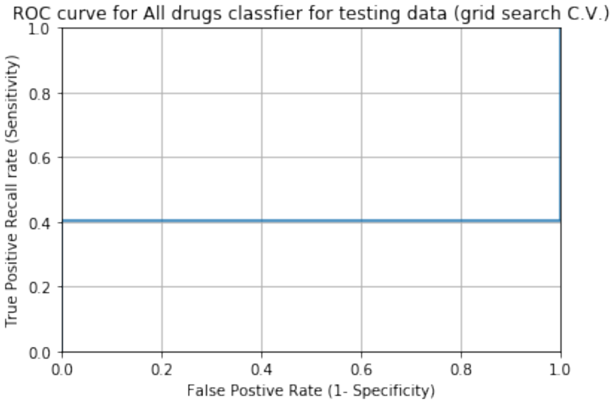
In summary, we expect Grid-search cross-validation to execute the best model performance followed by K-fold and train/test splitting.

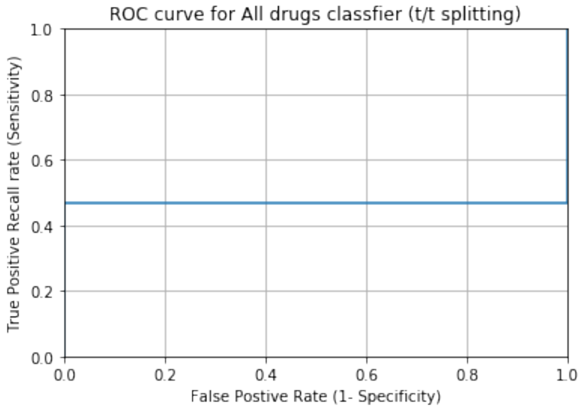
B. RESULTS FOR ALL DRUGS: ROC & ACCURACY


ROC curve for All drugs classfier for training data (k-fold C.V.)


ROC curve for All drugs classfier for testing data (k-fold with C.V.)


ROC curve for All drugs classfier for training data (Grid_Search C.V.)


ROC curve for All drugs classfier for testing data (grid search C.V.)

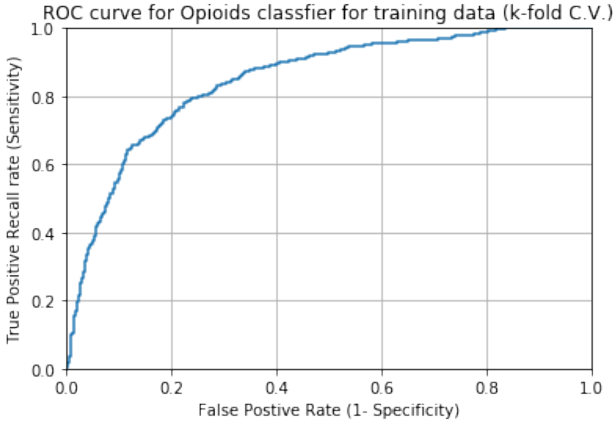


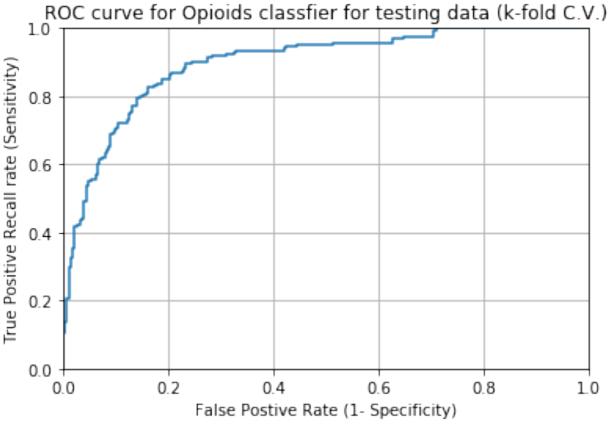
Feature	Accuracy (%)	Null Accuracy (%)	ROC Curve threshold 1(Sensitivity (%))	ROC Curve threshold 2(Specificity (%))	AUC score (%)
All drugs train/test splitting	99.79	99.79	100.00	0.00	50.00
All drugs train data w/ k-fold CV	99.86	99.86	100.00	0.00	50.00
All drugs test data w/ k-fold CV	99.79	99.79	100.00	0.00	50.00
All drugs train data w/ Grid-search CV	99.86	99.86	100.00	0.00	50.00
All drugs test data w/ Grid-search CV	99.79	99.79	100.00	0.00	50.00

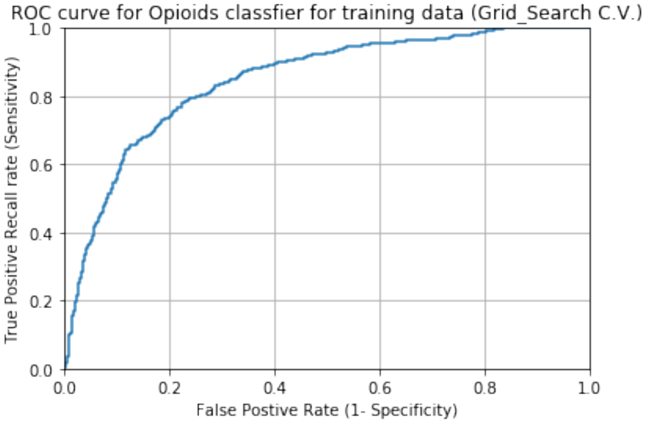
A ROC curve is the most commonly used way to visualize the performance of a binary classifier and the AUC score summarizes its performance in a single number by measuring its discrimination (the ability of the test to correctly classify those who use and don't use drugs). A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity).[12] In the evaluation of 1884 total respondents, 12 personality attributes served as dummy variables X and 18 psychoactive drugs served as dummy variables Y. Based on the results, we can conclude, that consumption for all drugs within a day/week/month/year is at least 99.97% accurate. The underlying distribution of the response values or types errors the classifier is making (Null accuracy) is at least 99.79%. Our ROC curve class distribution responses for all 5 models are suboptimal and provide poor results in predicting whether someone uses a drug or not.

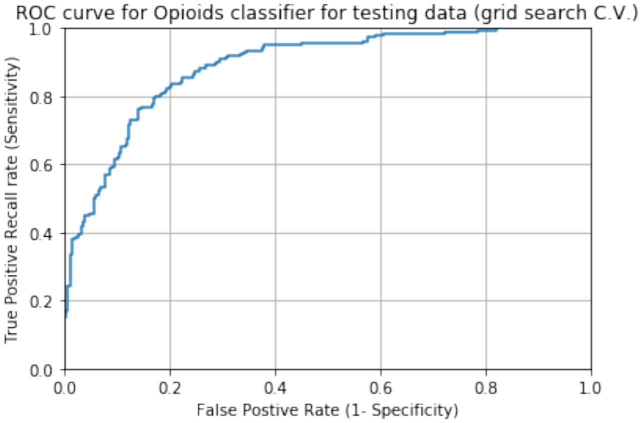
[12] Candy, J V, and E F Breitfeller. "Receiver Operating Characteristic (ROC) Curves: An Analysis Tool for Detection Performance." Aug. 2013, doi:10.2172/1093414.

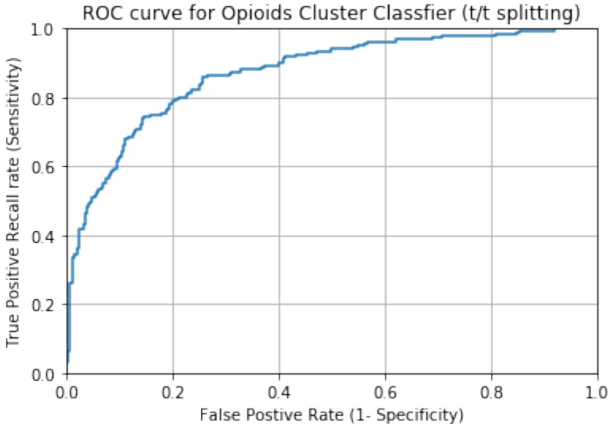
C. RESULTS FOR OPIOIDS CLUSTER: ROC & ACCURACY


ROC curve for Opioids classfier for training data (k-fold C.V.)


ROC curve for Opioids classfier for testing data (k-fold C.V.)


ROC curve for Opioids classfier for training data (Grid_Search C.V.)


ROC curve for Opioids classifier for testing data (grid search C.V.)


ROC curve for Opioids Cluster Classfier (t/t splitting)

We stated in the last section that the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test; positive discrimination. In the evaluation of 1884 total respondents, 12 personality attributes served as dummy variables X and the opioids cluster (heroin, meth, and legal highs) served as dummy variables Y. We conclude appropriate parameter/feature selection for all five models and state:

Feature	Accuracy (%)	Null Accuracy (%)	ROC Curve threshold 1(Sensitivity (%))	ROC Curve threshold 2(Specificity (%))	AUC score (%)
Opioids train/test splitting	80.68	60.93	72.28	86.06	79.17
Opioids Cluster train data w/ k-fold CV	78.87	63.58	66.28	85.95	76.12
Opioids Cluster test data w/ k-fold CV	82.59	60.93	75.54	87.10	81.33
Opioids Cluster train data w/ Grid-search CV	78.77	63.48	66.28	85.95	76.12
Opioids Cluster test data w/ Grid-search CV	80.89	60.93	67.39	87.80	78.96

1. *Accuracy:* The number of correct predictions made by all five models (for opioids use) over all kinds of predictions made is greater than 78.87%. The highest percent of correct classifications was obtained by the *k-fold test* model: 82.59%.
2. *Null Accuracy:* Provides information about the underlying distribution of the response values or types errors the classifier is making. The accuracy that could be achieved by always predicting the most frequent class for all five models is greater than 60.93%. The models that have the lowest percentage of iterating types of errors are all the testing systems.
3. *Sensitivity of ROC Curve:* Probability that a test result will be positive when a respondent uses an opioid (true positive). All five models have positive probabilities greater than 66.28% . *K-fold* test model had the highest probability threshold at 75.54%
4. *Specificity of ROC Curve:* Probability that a test result will be negative when a respondent doest not use an opioid (true negative). All five models have negative probabilities greater than 85.95% . *Grid-search test* model had the highest probability threshold at 87.80%
5. *AUC (area under the curve):* Measures discrimination, that is, the ability of the test to correctly classify those who use and don't use opioids. AUC can also be evaluated as a representation (if you choose one random and one negative observation) of the likelihood that the classifier will assign a higher predicted probability to the positive observation. Therefor, a higher AUC value is indicative of a better overall classifier and provides an alternative to the *Accuracy Score*. All five models provided very good scores for measuring discrimination greater than 76.12%. The *k-fold test* model had the highest *AUC Score* for correctly classifying respondents who use Opioids.

In summary, we are able to confirm the ability of our 5 models. All have demonstrated positive results when predicting the probability and discrimination of drug consumption for each individual based on personality attributes and the opioids cluster (heroin, methadone, and legal highs). In addition, the *Accuracy* and *AUC Scores* confirm proper model execution and tuning of the regularization parameter C. *Train/test splitting* provided the lowest results, followed by *Grid-Search* and *K-fold Cross-Validation.*

D. CLASSIFICATION REPORT FOR ALL DRUGS: PRECISION/RECALL/F-1

Feature	Drug User	Population	Precision	Recall	f-1 score
All drugs train/test splitting	NO [0]	1.0	0.0	0.0	0.0
	YES [1]	470.0	100.0	100.0	100.0
	avg / total	471.0	100.0	100.0	100.0
All drugs train data w/ k-fold CV	NO [0]	2.0	0.0	0.0	0.0
	YES [1]	1411.0	100.0	100.0	100.0
	avg / total	1413.0	100.0	100.0	100.0
All drugs test data w/ k-fold CV	NO [0]	1.0	0.0	0.0	0.0
	YES [1]	470.0	100.0	100.0	100.0
	avg / total	471.0	100.0	100.0	100.0
All drugs train data w/ Grid-search CV	NO [0]	2.0	0.0	0.0	0.0
	YES [1]	1411.0	100.0	100.0	100.0
	avg / total	1413.0	100.0	100.0	100.0
All drugs test data w/ Grid-search CV	NO [0]	1.0	0.0	0.0	0.0
	YES [1]	470.0	100.0	100.0	100.0
	avg / total	471.0	100.0	100.0	100.0

Classification Reports are generated using the fitted data with or without the models best regularization parameter C. It provides performance measurements in the form of proportion analysis (based on population). From the results of 1884 total respondents and their respective use of drugs (all), we can conclude that

1. The total train/test population of all models were correctly separated in a 3:1 ratio
2. Consumption of any drug was 100% probable.
3. For the five models, the proportion of respondents labeled as "using a drug" versus those who actually consumed it, was 100% precise (proportion performance false positives)
4. For the five models, we can state that every respondents case was labeled "User." The classifier's performance with respect to false negatives, provided a recall value of 100% for all test models. This is also confirmed with the f-1 score of 100%

The *Classification Reports* for all 5 models are suboptimal and provide poor proportion performance measurements.

E. CLASSIFICATION REPORT FOR OPIOIDS: PRECISION/RECALL/F-1

Feature	Drug User	Population	Precision	Recall	f-1 score
Opioids Cluster train/test splitting	NO [0]	287.0	83.0	86.0	84.0
	YES [1]	184.0	77.0	72.0	75.0
	avg / total	471.0	81.0	81.0	81.0
Opioids Cluster train data w/ k-fold CV	NO [0]	897.0	82.0	86.0	84.0
	YES [1]	516.0	73.0	66.0	70.0
	avg / total	1413.0	78.0	79.0	79.0
Opioids Cluster test data w/ k-fold CV	NO [0]	287.0	85.0	87.0	86.0
	YES [1]	184.0	79.0	76.0	77.0
	avg / total	471.0	82.0	83.0	83.0
Opioids Cluster train data w/ Grid-search CV	NO [0]	897.0	82.0	86.0	84.0
	YES [1]	516.0	73.0	66.0	70.0
	avg / total	1413.0	78.0	79.0	79.0
Opioids Cluster test data w/ Grid-search CV	NO [0]	287.0	82.0	88.0	85.0
	YES [1]	184.0	79.0	70.0	74.0
	avg / total	471.0	81.0	81.0	81.0

We stated in the last section that *Classification Reports* generate very important information about proportion performance measurements. From the results of 1884 total respondents and their respective use of opioids, we can conclude that:

1. *Precision:* A measure that tells us what proportion of respondents that were diagnosed as using an opioid, **actually** used the opioid. In other words, we are examining the precision of the models performance in respect to false positives. It is a common expectation for the *testing data* to have better precision than the *training data (*the test set error decreases with the increase of training set size). All five models have positive model precision greater than 73.00% (both in bias/variance) and consistent with the previous observation. *K-fold non-user test data* had the highest *precision* at 85.00% and averaged total result of 82.00%.
2. *Recall:* A sensitivity measure that tells us information about a classifier's performance with respect to false negatives (how many actual opioid users the model mislabeled as user/non-user). All five models have positive model recall greater than 66.00%. *Grid-Search non-user test data* had the highest performance *recall* at 88.00%. However, the highest averaged total result was for the *K-fold* model at 82.00%.
3. *F-1 Score:* The F-measure is the harmonic mean of the precision and recall scores. In most cases, you have a trade-off between precision and recall. If you optimize your classifier to increase one and disfavor the other, the harmonic mean quickly decreases. It is greatest however, when both precision and recall are equal. *K-fold non-user test data* had the highest *f-1 score* at 86.00% and averaged total result of 83.00%.

In summary, we are able to conclude a good overall performance ability for all of the 5 models; *Train/test splitting* provided the lowest performance, followed by *Grid-Search* and *K-fold Cross-Validation.* However, all demonstrated positive results when measuring *precision* of the actual use of an opioid, high *recall/sensitivity scores* outputting low percentages for user/non-user mislabeling, and good *f-1 scores* confirming the previous statements. Our findings are close in proportion with the referenced published paper. For the opioids pleiades, a *precision score* of 75.84% and *recall score* of 78.85% were reported. Therefore, we state that our best opioids model (*K-fold CV)* performed better than those found in the publication.

6. CONCLUSION

Implementation of appropriate exploratory/statistical and machine learning methods on selective data model testing has provided evidence in answering the client's problem. We analyzed the probability binary correlation of drug consumption for each individual based on personality attributes versus all psychoactive drugs and the Opioids cluster (heroin, methadone, and legal highs).

We report that the most vulnerable personality profile for substance abuse is a person that is an American white male, between the ages of 18-24, with some college education with high levels of neuroticism and openness to experience and dangerous levels of legal highs susceptibility, With 95% confidence, we also report that males are probable to the use of legal highs 71.45 ± 2.04% of the time, a result, that is consistently propositional to those found in the referenced publication.

Logistic regression models were built for exploration of possible binary correlation between personality features, all drugs and the opioids cluster. We report that the models used to evaluate personality traits and use of opioids have one, better performance measurements than those referenced, and second, positive results when predicting the probability and discrimination of opioid consumption. We also report that models used to evaluate personality traits versus use of all drugs were suboptimal, provide poor results in predicting whether someone uses a drug or not, and in essence, bad at performance measurements.

For future studies, we recommend a more in-depth model feature selection process. Logistic Regression may not be the most successful machine learning method when creating a framework for some of the proposed evaluations. By addressing possible overfitting, choice of parameters and/or imbalance of classes, it is possible to build a successful compatible system.

7. CLIENT RECOMMENDATIONS

Upon our findings, we recommend any current/future client to consider the following:

1. Addressing a more diverse surveyed pool of anonymous respondents (and methods of acquiring the data) is vital in providing more precise reporting. A 2016 publication by the Centers for Disease Control and Prevention[13], concluded that minorities who have medical history regarding substance abuse, often don't report it in Census and therefore become ghosts in data surveys such as the one utilized in this review. We found evidence of this in the ethnicity and country exploratory sections.
2. A recent Chicago tribune article discovered that fentanyl, a dangerous chemical compound found in synthetic opioids (part of the legal highs pleiades) is behind the death increase in 20 states, with 10 doubling their rates from 2015 to 2016.[14] These findings are consistent with our ideal personality profile for substance abuse. We recommend allocation of money to be centered towards two areas:
    i. Drug prevention centers and appropriate programs in places of higher education. For those with behavioral susceptibility, regardless of gender, college appears to be a foundation for substance abuse.
    ii. Increase in legal high awareness. According to our findings, the root to this problem appears to happen when the individual falls in the age bracket of 18-24-which we can deduce of earlier age experimentation.

[13] "Increases in Drug and Opioid Overdose Deaths - United States, 2000–2014." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 1 Jan. 2016, www.cdc.gov/mmwr/preview/mmwrhtml/mm6450a3.htm.

[14] Candy, J V, and E F Breitfeller. "Receiver Operating Characteristic (ROC) Curves: An Analysis Tool for Detection Performance." Aug. 2013, doi:10.2172/1093414.