bda_project

by Sean Shiverick

Submission date: 03-Dec-2017 11:47PM (UTC-0500)
Submission ID: 889566101
File name: bda_project.pdf (929.19K)
Word count: 8739
Character count: 47690

Using Machine Learning Classification of Opioid Addiction for Big Data Health Analytics

Sean M. Shiverick
Indiana University Bloomington
smshiver@indiana.edu

ABSTRACT

Classification of opioid addiction can identify important features relevant for predicting drug abuse and overdose death. Machine learning procedures were applied to data from a large National Survey of Drug Use and Health (NSDUH-2015) to classify individuals for illicit opioid use according to demographic characteristics and mental health attributes (e.g., depression). Classification models of opioid addiction can be extended for big data health analytics to include high-dimensional datasets, data collected over previous years, or expanded to the larger population of patients taking prescription opioid medication. The results seek to raise awareness of risk factors related to opioid addiction among patients and medication prescribers, and help decrease the risk of opioid overdose death.

KEYWORDS

Health Analytics, Machine Learning Classifiers, Opioid Addiction, Big Data, i523, hid335

1 INTRODUCTION

Big Data offers tremendous potential to fuel innovation and transform society. Can this momentum be harnessed to address a serious health crisis such as the opioid overdose epidemic? [7] Health informatics is generating huge amounts of data at a rapid pace, from electronic medical records (EMRs), clinical research data, to population-level public health data [5]. This project considers health analytics from two levels, the research questions being addressed and the data used to answer them. The question of interest in this project is whether opioid dependency and addiction can be predicted from demographic attributes and psychological characteristics. Survey research provides data on a wide range of issues that people may be reluctant to disclose, including mental health disorders, personal medical health concerns, prescription medications, and illicit drug use. Responses to surveys may be biased to some degree, but measures of confidentiality and anonymity help to assure more accurate disclosures. The goal of this project is to use machine learning procedures to classify individuals susceptible to opioid abuse and dependence. Understanding the features that contribute to opioid addiction can identify underlying risk factors and increase awareness of potential opioid abuse for patients and health care providers. The results could be extended to big data from previous years of the opioid crisis and to the larger population of patients taking prescription opioid mediation. Different machine learning classification methods are discussed.

1.1 Opioid Overdose Epidemic

The abuse of prescription opioid medication in the U.S. has become a major health crisis of epidemic proportions [23]. Over 2 million Americans were dependent or abused prescription opioids such as oxycodone or hydrocodone in 2014[3]. Overdose deaths from prescription opioids have quadrupled since 1999, resulting in more than 180,000 deaths between 1999 to 2015 [11]. Drug overdose deaths increased significantly for males and females, between 25-44 years, ages 55 and older, for Non-Hispanic Whites and Blacks, in the Northeast, Midwest, and Southern regions of the U.S. [7]. Mobile health applications can monitor patient medication consumption and provide an early warning system for potential abuse, detecting sudden changes in medications, higher dosages, or rapid escalation of a prescribed dosage [22]. Reliable information about medication dosages can be difficult to obtain based on self-reports. Individuals dependent or addicted to prescription opioids may obtain synthetic opioids such as fentanyl or illicit drugs such as heroin. Because the dosage levels and potency of illicit opioids are largely unknown, there is greater risk of drug overdose death. The sharp increase in overdose deaths due to synthetic opioids (other than methadone) has coincided with the increased availability of illicitly manufactured fentanyl, which is indistinguishable from prescription fentanyl. The findings indicate the opioid overdose epidemic is getting worse, and requires urgent action to prevent opioid dependence, abuse and overdose death. The target group for this project is individuals who reported misusing or abusing prescribed opioid medication who also used heroin, shown in Figure 1.

1.2 Machine Learning Approaches

Machine learning is a set of procedures and automated processes for extracting knowledge from data. The two main branches of machine learning are supervised learning and unsupervised learning. Supervised learning problems involve prediction about a specific target variable or outcome of interest. If a given dataset has no target outcome, unsupervised learning methods can be used to discover underlying structure in unlabeled data. The goal of this project is to classify opioid addiction and focuses on supervised learning. Supervised learning is used to predict a certain outcome from a given input, when examples of input/output pairs are available [10]. A machine learning model is constructed from the training set of input-output pairs, to predict new test data not previously seen by the model. The two major approaches to supervised learning problems are regression and classification. When the target variable to be predicted is continuous, or there is continuity between the outcome (e.g., home values, or income), a regression model is used to test the set of features that predict the target variable. If the target is a class label, set of categorical or binary outcomes (e.g., 'spam' or 'ham', 'benign' or 'malignant'), then classification is used

to predict which class or category label that new instances will be assigned to.

1.3 Classification Algorithms

Comparing the performance of different learning algorithms can be helpful for selecting the best model for a given problem [14]. One of the simplest classification algorithms is K-Nearest-Neighbors (KNN) which takes a set of data points and classifies a new data point based on the distance (e.g., Euclidean, by default) to its nearest neighbors. The main parameter for KNN is the number of neighbors, and k of 3 or 5 neighbors works well. The advantage of the KNN classifier is that it provides a solution that is easy to understand. A limitation of KNN is that it does not perform well with a large number of features (100 or more) or sparse datasets. Several different classification algorithms are considered below.

1.3.1 Logistic Regression Classifier. Logistic regression is a commonly used linear model for classification problems. The decision boundary for the logistic regression classifier is a linear function of the input; a binary classifier separates two classes using along a line, plane, or hyperplane. Linear classification models differ in terms of (1) how they measure how well a particular combination of coefficients and intercept fit the training data, and (2) the type of regularization used [10]. The main parameter for linear classification models is the regularization parameter 'C'. High values of C correspond to less regularization and the model will fit the training set as best as possible, stressing the importance of each individual data point to be classified correctly. By contrast, with low values of C, the model puts more emphasis on finding coefficient vectors (w) that are close to zero, trying to adjust to the 'majority' of data points [10]. In addition, the penalty parameter influences the coefficient values of the linear model. The L2 penalty (Ridge) uses all available features, but pushes the coefficient values toward zero. The L1 penalty (Lasso) sets the coefficient values for most features to zero, and uses only a subset for improved interpretability. This paper uses a logistic regression classifier to predict Heroin use from demographic attributes, mental health, prescription opioids, medication use, misuse, and illicit drug use.

1.3.2 Tree Based Models. Decision tree models are widely used for classification and regression. Tree models "learn" a hierarchy of if-else questions that are represented in the form of a decision tree. Building decision trees proceeds from a root node as the starting point and continues through a series of decisions or choices. Each node in the tree either represents either a question or a terminal node (i.e.,leaf) that contains the outcome. Applied to a binary classification task, the decision tree algorithm "learns" the sequence of if-else questions that arrives at the outcome most quickly. For data with continuous features, the decisions are expressed in the form of, "Is feature x larger than value y?" [10] In constructing the tree, the algorithm searches through all possible decisions or tests, and find a solution that is most informative about the target outcome. A decision tree classifier is used for binary or categorical targets, and decision tree regression is used for continuous target outcomes. The recursive branching process of tree based models yields a binary tree of decisions, with each node representing a test that considers a single feature. This process of recursive partitioning is repeated until each leaf in the decision tree contains only a single target. Prediction for a new data point proceeds by checking which region of the partition the point falls in, and predicting the majority in that feature space. The main advantage of tree based models is that they require little adjustment and are easy to interpret. A drawback is that they can lead to very complex models that are highly overfit to the training data. A common strategy to prevent overfitting is *pre-pruning*, which stops tree construction early by limiting the maximum depth of the tree, or the maximum number of leaves. One can also set the minimum number of points in a node required for splitting. Another approach is to build the tree and then remove or collapse nodes with little information, which is called *post-pruning*. Decision trees work well with features measured on very different scales, or with data that has a mix of binary and continuous features.

1.3.3 Random Forests Classifier. A random forest is a collection of decision trees that are slightly different from the others, which each overfits the data in different ways. The idea behind random forests is that overfitting can be reduced by building many trees and averaging their results. This approach retains the predictive power of trees while reducing overfitting. Randomness is introduced into the tree building process in two ways: (a) selecting a bootstrap sample of the data, and (b) selecting features in each node branch [10, 14]. In building the random forest, we first decide how many trees to build (e.g., 10 or 100), and the algorithm makes different ransom choices so that each tree is distinct. The bootstrapping method repeatedly draws random samples of size n from the dataset (with replacement). The decision trees are build on these random samples that are the same size as the original data, with some points missing and some data points repeated. The algorithm also selects a random subset of p features, repeated separately each node in the tree, so that each decision at the node branch is made using a different subset of features. These two processes help ensure that all of the decision trees in the random forest are different. The important parameters for the random forests algorithm are the number of sampled data points and the maximum number of features; the algorithm could look at all of the features in the dataset or a limited number. A high value for "maximum features" will produce trees in the random forest that are very similar and will fit the data easily based on the most distinctive features, whereas a low value will produce trees that are very different from each other, and reduces over- fitting. Random forests is of the most widely used ML algorithms that works well without very much parameter tuning or scaling of data. A limitation of this approach is that Random forests do not perform well with very high- dimensional, data that is sparse data, such as text data.

1.4 Project Goals

The general idea of the project is that prescription opioid dependency and addicted will in many cases lead to the use of illicit opioids such as heroin or fentanyl. According to this reasoning, it was hypothesized that individuals who report using heroin may also be susceptible to misusing or abusing prescription opioid medications. The goal of the study was to identify the set of features important for predicting opioid addiction. The data used in the project is from the National Survey on Drug Use and Health from 2015 (NSHUH-

2015) [1], which is the most recent year available. The NSDUH-2015 is a comprehensive survey that covers all aspects of substance use, misuse, dependency, and abuse, including questions related to both prescription medications (opioids, tranquilizers, sedatives) and illicit drugs (e.g., heroin, cocaine, methamphetamine), drug dependency, addiction, and treatment, demographic measures of education and employment, physical health, depression, and mental health treatment. Several classification models were constructed to classify heroin use in the sample by demographics attributes and mental health characteristics (e.g., adult depression). This method addresses the following issues related to opioid dependency and addiction: (i) Identify factors related to illicit opioid use, (ii) Identify factors related to prescription opioid misuse and abuse, and (iii) Examine the relationship between prescription opioid misuse, abuse and heroin use.

2 METHOD

2.1 Data

Data from the 2015 NSHUH was downloaded from the Substance Abuse and Mental Health Data Archive (SAMHDA) [1] URL using the 'get-data.py' function written to unzip the data files, extract the data as a Pandas data frame, and write the file to CSV file [4]. The dataset consists of 57,146 observations with 2,666 features representing individual-level responses from a survey of the U.S. population. According to the NSDUH codebook, sampling was weighted across states by population size for a representative distribution selected from 6,000 area segments. The sample design used five state sample size groups drawing more heavily from the eight states with the largest population (e.g., CA, FL, IL, MI, NY, OH, PA, TX) which together account for 48 percent of total U.S. population aged 12 or older. All identifying information was collapsed (e.g., age categories) and state identifiers were removed from the public use file to ensure confidentiality. The NSDUH public-use files do not include geographic location, or demographic variables related to ethnicity or immigration status. The weighted survey screening response rate was 81.94 percent and the weighted interview response rate was 71.2 percent.

2.2 Data Cleaning and Preparation

2.2.1 Data Cleaning. All steps of this analysis was completed in a python interactive notebook [16] based following examples from *Python for Data Analysis* [9]. After saving the NSDUH-2015 as a data frame object, the dataset was subset by columns to include demographic characteristics (e.g., age category, sex, marital status, education, employment status, and category of metropolitan area), measures of physical health (e.g., overall health, STDs, Hepatitis, HIV, Cancer, hospitalization), mental health (e.g., Adult Depression, Emotional Distress, Suicidal Thoughts, Plans), Suicide Attempts, Pain Reliever Medication Use, Misuse, and Abuse (over past year, past month), Prescription Opioid Medications Taken in Past year (e.g., Hydrocodone, Oxycodone, Tramadol, Morphine, Fentanyl, Oxymorphone, Demerol, Hydromorphone), Heroin Use, Abuse (over past year, past month), Tranquilizer Use, Sedative Use, Cocaine Use, Amphetamine and Methamphetamine Use, Hallucinogen Use, Drug Treatment (e.g., Inpatient, Outpatient, Hospital, Mental Health Clinic, ER, Drug Treatment Status), and Mental Health Treatment History. A codebook was created to provide a complete list of variables included with summaries of response categories [19]. The following steps were taken to detect and remove inconsistencies in the data [13]:

(1) Remove missing values (i.e., 'NaN')
(2) Recode blanks, non-responses, or legitimate skips (e.g.,'99', '991', '993') to zero
(3) Recode dichotomous responses (e.g., "Yes=1"/"No=2") so that "No=0"
(4) Recode categorical variables to be consistent with amount or degree (e.g., "1=low", "2=med", "3=high")
(5) Rename selected variables for better description (e.g., Adult Major Depressive Episode Lifetime changed from 'AMDELT' to 'DEPMELT')

2.2.2 Aggregate Variables. Because the majority of features were represented as dichotomous "Yes/No" variables, related features were summed to create aggregated variables. For example, overall health, STD, Hepatitis, HIV, Cancer, and hospitalization were aggregated to create a single health measure. The health measure was recoded so that higher scores indicated better health. Questions related to depression, emotional distress, and suicidal thoughts were summed to create a single variable for mental health ('MENTHLTH') with scores ranging from 0 to 9. Responses to pain reliever medication use, misuse, abuse, or dependency, were aggregated to create a single variable of pain reliever misuse or abuse ('PRLMISAB'). All prescription painkiller medications used in the past year were summed. Similarly, all related responses were summed to create single variables for Tranquilizers, Sedatives, Cocaine, Amphetamines, Hallucinogens, Drug Treatment, and Mental Health Treatment. The target outcome of interest for classification, lifetime heroin use (i.e., "Have you ever used heroin before, at any time?") is a dichotomous variables. The demographic characteristics and aggregated variables were subset and saved to a new data frame consisting of 2 features and 57,146 observations, which was exported to CSV file.

3 RESULTS

3.1 Exploratory Data Analysis

Of the total sample of N=57,146 respondents, 26,736 were male and 30,410 female; 6,343 individuals reported misusing pain medication at some point; however, only 956 repondents had used heroin (570 males, 386 females). Table 1 shows the raw counts of individual substance use by age group (with the sample size for each age group), listing the ten most commonly used opioid pain medications, self-reported misuse of prescription opioid pain relievers (i.e., PRL Misuse Ever), use of prescription Tranquilizers, Sedatives, and Methadone. In addition, self-reported use of illicit drugs such as heroin, cocaine, amphetamines, methamphetamine, Hallucinogens, including LSD and Ecstasy (MDMA). This summary table shows that substance use seems to be highest for individuals between the ages of 18 to 25 and from 35 to 49 years. Of the prescription relievers, Hydrocodone use was almost double the rate of Oxycodone use for each age group, and was significantly higher than any other prescription opioid medication. Use of prescription Fentanyl and

Demerol, two powerful opioids, and synthetic morphines such as Oxymorphone and Hydromorphone, was very low. The rate of prescription Tranquilizer use was several orders of magnitude higher than Sedative use or Methadone use. Compared to other illicit drugs such as Cocaine, Amphetamines, Hallucinogens, heroin use was not very common in this sample. The highest rates of heroin use were seen between the ages of 18 to 49, and was lowest for respondents in the youngest age group 12 to 17, and individuals over 50.

[Table 1 about here.]

Table 2 shows the frequency of individuals reporting that they had experienced mental health issues such as depression, suicidal thoughts,whether they had received mental health treatment, received treatment from a private therapist, or believed that they needed drug treatment, but had not sought treatment, across each age category. Frequency of depression was not included for respondents between 12 to 17 years, and the measure was of adult depression.

[Table 2 about here.]

Figure 1 shows the proportion of individuals who reported misusing prescription opioid pain relievers and who reported using heroin. The left column of the Figure 1 shows the majority of respondents (89 percent) stated they had never misused prescription opioid pain medication or used heroin, although 10 percent reported misusing opioid pain medication at some point. The right panel of Figure 1 shows that, of those individuals who reported using heroin, the proportion who also reported misusing opioid pain medication was almost twice as large as the proportion of those who only used heroin. This is consistent with the hypothesis that misuse of prescription opioids is linked with heroin use for some individuals.

[Figure 1 about here.]

Figure 2 shows the aggregated measure of Opioid Pain Reliever misuse and abuse plotted against the aggregated measure of Heroin use (which includes misuse, abuse, lifetime use, past year use, 30 day use), with weighted regression lines grouped by size of City Metropolitan region (from none to large). The largest proportion of the sample who report prescription opioid misuse, abuse, and heroin use is represented by observations from large metropolitan areas (red circles) with large population size. However, a small number of observations from rural or small metropolitan regions (blue and green circles) showed very high rates of prescription opioid misuse and abuse. Regression lines (i.e., line of best fit) shown are weighted by the City/Metro region attribute, with a steeper slope shown for smaller metropolitan regions than large metropolitan regions. The difference in slope may be due to the influence of the small number of outliers who had high degrees of prescription opioid misuse, and heroin use. The plot also shows a clear divide on the y-axis, which separates the sample according to high and low or no prescription opioid misuse, although the continuum of heroin use from no, low, to high is distributed fairly evenly along the x-axis.

[Figure 2 about here.]

Figure 3 shows the pairplots of demographic features including mental health (higher scores equal to more depression), Prescription Opioid Pain Reliever (PRL) Medication (aggregated), Heroin Use (aggregated measure), and Size of City/Metropolitan region. The top row shows that the majority of the sample reported no mental health concerns, whereas a small proportion of the sample reported depression, emotional distress, or suicidal thoughts. Only few people self-described as high in depression reported low Prescription Opioid PRL misuse and abuse. The plot also reveals that prescription opioid misuse and heroin use were distributed approximately evenly for individuals reporting either low, moderate, or high levels of depression, which suggests that depression was not a factor in predicting opioid misuse. The second row shows a small number of individuals from rural areas or small cities who reported very high levels of prescription opioid misuse, although the majority of respondents misusing or abusing prescription opioid were from large metropolitan areas. As described above, the majority of respondents (about 90 percent of the sample) reported they had never misused prescription opioids. In the second row and third and fourth columns, a natural break is seen between individuals who reported high levels of prescription opioid misuse and abuse and those who reported very low or no opioid misuse. A very small proportion of the entire sample reported both misusing and abusing prescription opioids and using heroin, but this is a group of interest. The last column of the second row shows the individuals reporting high levels of opioid misuse and abuse were distributed evenly across city/metropolitan areas of different sizes, with only slightly higher numbers for small cities or rural areas. As stated above, only few participants reported using heroin, and of these, the majority were from large metropolitan areas. Finally, the sample seems to have slightly higher proportions from small and large metropolitan areas, which is likely due to weighted sampling, which drew more from heavily populated regions.

[Figure 3 about here.]

3.2 Classifier Models of Heroin Use

This analysis classified individuals according to whether they had ever used heroin (i.e., "Heroin Use Ever"). All classifier models were constructed using SciKit Learn [10] using an interactive python jupyter notebook [17]. The features of interest were demographic characteristics, health, mental health (adultdepression), prescription opioid misuse and abuse ('PRLMISEVR', 'PRLMISAB', 'PRLANY'), prescription tranquilizers and sedatives ('TRQLZRS', 'SEDATVS'), illicit drugs ('COCAINE', 'AMPHETMN'), drug treatment ('TRTMENT'), and mental health treatment ('MHTRTMT'). The target variable was Heroin Use ('HEROINEVR'). Next, the dataset was split into the training set and test sets using the 'train-test -split' function in 'sklearn'. Model accuracy for the training set and test set are reported, with different parameter values, and features importance.

3.2.1 Logistic Regression Classifier. Logistic Regression Classification is a based on a linear equation that calculates the relative weight of each feature for a categorical target or binary outcome ("yes/no")[14]. The logistic regression classifier was fit to the training data in Scikit-Learn, and the model was validated on the test data. By default, the model applies L2 penalty (Ridge). The training set accuracy was 0.983 and the test set accuracy was 0.984. The parameter 'C' determines the strength of regularization, with higher values of C providing greater regularization. The L1 penalty

(Lasso) limits the values of most coefficients to zero, creating a more interpretable model that uses only a few features. Figure 4 plots the coefficients of logistic regression classifier for heroin use with the L1 Penalty (Lasso) under different values of parameter C. The default setting, C=1.0, provides good performance for train and test sets, but the model is very likely underfitting the test data. Using a higher value of C fits a more 'flexible' model and generally gives improved accuracy for both training and tests sets. Using a value of C=100 yielded training set accuracy of 0.98 and test set accuracy of 0.98. Figure 4 shows that the features coefficient values did not change much according to the values of parameter C, and the accuracy values were approximately the same for all values of C. Examination of the coefficients from the logistic regression classifier revealed the three features which were most closely associated with Heroin use were: Prescription Opioid Pain Reliever (PRL) Misuse ever (as predicted), Cocaine Use, and Amphetamine use, respectively.

[Figure 4 about here.]

3.2.2 Decision Tree Classifier. The following analysis used the *Decision Tree Classifier* package in Scikit-Learn, which only does pre-pruning. First, the decision model was build using the default setting of a fully developed tree until all leaves are pure. The random state' features is fixed to break ties internally. Accuracy on the training set was 0.99 and test set accuracy was 0.974. Without restricting their depth, decision trees can become complex; unpruned trees are prone to overfitting and do not generalize well to new data. Limiting the depth of tree decreases overfitting, which results in lower training set accuracy, but improved performance on the test set. Next, pre-pruning was applied, with a maximum depth of 4, which means the algorithm split on four consecutive questions. Training set accuracy of the pruned tree was 0.985 and test set accuracy was 0.984. Even with a depth of 4, the tree can become a bit complex. Figure 5 shows a partial view of the decision tree classifier of heroin use (the entire tree was too wide to include as a legible Figure), and the full tree image is available in the notebook 'BDA-Analytics-Classifier-Heroin.ipynb' [17]. The decision tree shows the top features that the algorithm split on to classify heroin use. One way to interpret a decision tree it by following the sample numbers represented at the test split for each node. The classifier algorithm selected Cocaine Use (aggregated score) as the root node of the decision tree. The branch to the left side of the tree represents samples with a score equal to or less than 1.5 (n=40956), whereas the branch to the right represents samples with a Cocaine Use score greater than 1.5 (n=1903). The second split on the right occurs for Any Prescription Opioid Pain Reliever Use ('PRLANY'), with n=1443 having a score less than or equal to 3.5, and n=460 respondents with a PRL score greater than 3.5. In other words, of those respondents who reported relatively high Cocaine use, a small portion also reported relatively high Prescription Opioid PRL use. Instead of looking at the whole tree, features importance is a common summary function that rates how important each feature is for the classification decisions made in the algorithm. Each feature is assigned an importance value between 0 and 1; with a value of 1 indicating the feature perfectly predicts the target and a value of 0 meaning that the feature was not used at all. Feature importance values also always sum to 1. A feature may have a low feature importance value because another feature encodes the same information. The top two important features for classifying Heroin Use were Cocaine Use and Any Prescription Opioid PRL Use, with smaller importance given to Opioid PRL Misuse Ever and Prescription Opioid PRL Misuse and Abuse.

[Figure 5 about here.]

3.2.3 Random Forests Classifier. Random forests is an ensemble approach that builds many trees and averages their results to reduce overfitting. The model was build using the Random Forest Classifier package in Scikit-Learn. The parameters of interest for builiding random forests are: (a) the number of trees ('n-estimators'), (b) the number of data points for bootstrap sampling ('n-samples'), and (c) the maximum number of features considered at each node ('max-features'). The max-features parameter determines how random each tree is, with smaller values of max-features resulting in trees in the random forest that are very different from each other. This analysis applied a random forest consisting of 100 trees to classify Heroin Use, and the random state was set to zero. The training set accuracy was 0.999 and the test set accuracy was 0.984. Often the default settings for random forests work well, but we can apply pre-pruning as with a single tree, or adjust the maximum number of features. Feature importance for random forests is computed by aggregating the feature importance over trees in the random forest, and random forests gives non-zero importance to more features than a single tree. Typically random forests provide a more reliable measure of feature importance than the feature importance for a single tree. Figure 6 shows the feature importance of the random forests classifier for heroin use with 100 trees. Similar to the single tree, the random forest selected Cocaine Use as the most informative feature in the model, followed by Any PRL Use, which is an aggregated measure of prescription opioid medication use. Following after that, several features were tied for third place of importance, namely Education Level, Overall Health, Age Category, and Pain Reliever Misuse and Abuse. Random forests provides much of the same benefit as decision trees, while compensating for some of their shortcomings of overfitting. Single trees are still useful for visually representing the decision process.

[Figure 6 about here.]

3.2.4 Gradient Boosting Classifier Tree. Gradient boosting machines is another ensemble method that combines multiple decision trees for regression or classification by building trees in a serial fashion, where each tree tries to correct for mistakes of the previous one [10]. Gradient boosted regression trees use strong pre-pruning, with shallow trees of a depth of one to five. Each tree only provides a good estimate of part of the data, but combining many shallow trees (i.e., "weak learners"), the use many simple models iteratively improves performance. In addition to pre-pruning and the number of trees, an important parameter for gradient boosting is the learning rate, which determines how strongly each tree tries to correct for mistakes of previous trees. A high learning rate produces stronger corrections, allowing for more complex models. Adding more trees to the ensemble also increases model complexity. Gradient boosting and random forests perform well on similar tasks and data; it is common to first try random forests and then include gradient boosting to attain improvements in accuracy of the learning

model. This analysis used the Gradient Boosting Classifier from Scikit-Learn to classify Heroin Use, with the default setting of 100 trees of maximum depth of 3, and a learning rate of 0.1. The model was build on the training set and evaluated on the test set, with both training set and test set accuracy equal to 0.984. To reduce overfitting, pre-pruning could be implemented by reducing the maximum depth, or by reducing the learning rate. Figure 7 shows that the feature importance for the gradient boosting classifier tree looks similar to the feature importance for random forests, but the gradient boosting has decreased the importance of many features to zero. Again Cocaine is selected as the most imformative features, followed by Any Opioid PRL Use. In addition to Prescription Opioid PRL Misuse and Abuse, the gradient boosting classifier selected Amphetamine Use as an informative feature of Heroin Use.

[Figure 7 about here.]

3.3 Classifier Models of Prescription Opioid Pain Reliever (PRL) Misuse

This section reports results from the same set of classification analyses described above using Prescription Opioid Pain Reliever Misuse ('PRLMISEVR') the target variable. Attributes related to Heroin Use were now included as features (e.g., 'HEROINEVR', 'HEROINUSE', 'HEROINFQY'). The classifier models were built using SciKit Learn in a python notebook [18]. The dataset was split into the training set and test sets using the 'train-test- split' function in sklearn and the target variables was designated. Model accuracy for the training set and test set are reported, for different parameter values, with feature importance.

3.3.1 Logistic Regression Classifier. The logistic regression classifier was fit to the training data using the L1 penalty (Lasso), using different values of the regularization parameter C, and the model was validated on the test data. Higher value of parameter C typically gives improved accuracy for both training and tests sets; however, in this case, the training set accuracy was 0.901 and test set accuracy was 0.903, and these values were consistent for all values of parameter C. Figure 8 plots the coefficients of logistic regression classifier for Prescription Opioid PRL Misuse under different values of C. As shown in Figure 8, the features with the highest coefficient values were Treatment (for substance use), Heroin Use (as predicted), as well as Cocaine and Amphetamine use. This result indicates that Prescription Opioid Misuse is positively related to Drug Treatment, meaning that respondents who reported higher levels of opioids misuse were also in treatment, but that people who were misusing opioid medications were also more likely to have used illicit drugs such as heroin, cocaine, and amphetamine.

[Figure 8 about here.]

3.3.2 Decision Tree Classifier. The Decision Tree Classifier package in Scikit-Learn was used to build the tree model, pre-pruning was applied with a maximum depth of 4, which means the algorithm split on four consecutive questions. The training set accuracy of the pruned tree was 0.902 and test set accuracy was 0.902. Figure 9 shows a partial view of the decision tree classifier of prescription opioid misuse (the full tree is included in the 'BDA-Analytics-Classifier-PRL.ipynb' notebook) [18]. As Figure 9 shows, the decicion tree classifier selected Cocaine Use as the root note, that branched by the test score equal to or less than 0.5 (any Cocaine Use). At the second node, on the branch to the right n=5015 samples were further divided according to heroin use, with n=1913 having a score greater than 0.5 (any Heroin Use). At the third node on the right branch, samples were selected according to Tranquilizer medication use, with n=1419 scoring positively. On the left branch, the second node selected was Drug Treatment, with n=2844 respondents scoring positively that they had received Drug Treatment. Feature importance of the decision tree classifier selected Cocaine Use as the most informative feature for Prescription Opioid PRL Misuse. Following afterwards, Tranquilizer Use, Drug Treatment, and Heroin Use were tied for second place.

[Figure 9 about here.]

3.3.3 Random Forests Classifier. The Random Forest Classifier package in Scikit-Learn was used to classify Prescription Opioid PRL Misuse as the target variable, with 100 trees. The model accuracy for the training set was 0.955 and the test set accuracy was 0.896, which suggests that the model overfit the data. Figure 10 shows the feature importance of the random forests classifier for Prescription Opioid PRL Misuse. As Figure 10 shows, several features were identified as important for classifying Prescription Opioid PRL Misuse. The random forest selected Overall Health as the most informative feature in the model, followed by Cocaine Use, Education Level, Age Category, and Size of City Metropolitan region. Because of the additional features included as important, gradient boosting was performed to clarify the feature importance.

[Figure 10 about here.]

3.3.4 Boosted Gradient Classifier. The Gradient Boosting Classifier from Scikit-Learn was used to classify Prescription Opioid PRL Misuse, using the default setting of 100 trees, of maximum depth of 3, and a learning rate of 0.1. The model accuracy for the training set was 0.894 and accuracy for the test set was 0.893. Gradient boosting typically improves test set accuracy by using many simple models iteratively. In this case, model accuracy for gradient boosting was no better than random forests, and this is because the default parameter settings were used; further parameter tuning is needed to improve model performance. Feature importance was a primary interest for identifying features related to ' prescription opioid abuse. Figure 11 shows the feature importance for the gradient boosting classifier tree. As Figure 11 shows, several features were important for classifying prescription opioid misuse, and contrary to the random forests, gradient boosting selected Tranquilizer use as the most informative feature. Following closely in importance were Heroin Use and Age Category. Tied for fourth place were Cocaine Use and Treatment, with Mental Health (depression) coming in fourth in terms of feature importance. This result illustrates that several features are important for understanding Prescription Opioid Misuse, and the relations among features may be complex.

[Figure 11 about here.]

4 DISCUSSION

4.1 Summary of Findings

4.2 Extension to Big Data

The methods used in this project could be extended to better approximate big data for health analytics in the following ways: (1) Include a larger selection of features from the 2600 attributes in in the NSDUH-2015 dataset; (2) Include survey data from previous years (e.g., 2005-2015); and (3) Extend the sample to the population of patients who have been prescribed opioid pain medication. There were many additional features that could have been included in the subset of features included in the project dataset. However, data cleaning and preparation can be a time consuming process, especially for datasets with a large number of features [13]. Additional data from the NSDUH was downloaded from previous years (2012 to 2014), and a preliminary examination of the data revealed inconsistencies in questions and prescription opioid medications that would need to be resolved in order to combine data from multiple years. In addition to data cleaning, there are several steps involved in the consolidation of data from multiple sources into a single dataset, which include extraction, integration, and aggregation of features. Unfortunately, time constraints for the project deadline did not allow for the inclusion of data from previous years into this analysis. A future study could integrate data from different years into the analysis or include data from multiple sources.

4.3 Limitations

To be of any use, diverse and often messy raw data has to be sifted through and effectively organized for further analysis, and

there are legitimate questions about the reliability of self report data from surveyresearch for predicting actual behavior.

The question of Value evaluates the quality of the data as it pertains to intended outcomes, such as limiting the spread of contagion and disease prevention.

An important challenge for making sense of big data is developing analytic tools adequate to handle large volumes of data in real time.

4.4 Drug Abuse, Dependency, and Addiction

Drug addiction has many similar characteristics to other chronic medical illnesses, but there are unique challenges to the treatment of addiction [8, 20]. In drug rehabilitation treatment programs, patients undergo intense detoxification that reduces their drug tolerance, but are then released back into the environments associated with their drug use, putting them at high risk for relapse and potential drug overdose [6]. According to a classical conditioning model of addiction, situational cues or events can elicit a motivational state underlying relapse to drug use. Addictive behavior can be also be reinstated after extinction of dependency by exposure to drug-related cues or stressors in the environment [15].

4.5 Dynamics of Epidemic Spreading

If the prescription opioid crisis is a guenuine epidemic, then we can conceive of it in terms of the dynamics of epidemic spreading which have been developed based on models of contagious disease. Epidemic spreading is a dynamic process based on networks of direct person-to-person contact and indirect exposure via transportation pathways [2], that facilitate the distribution of opioid medications or illicit drugs. Instead of thinking about persons as infected or uninfected by biological contagion, in the opioid drug model, we must consider individuals as dependent, addicted or susceptible to dependence and addiction. Epidemics are quantified in terms of the proportion of the population infected, those yet to be infected, and the rate of transmission. Furthermore, the structure of the contact network can influence epidemic spreading [12]. For example, in the case of simple contagion, weak ties among acquaintances or infrequent associations provide shortcuts between distant nodes that reduce distance within the network [?] which can facilitate the spread of contagion, or in this case drug use. Furthermore, opioid contact networks may have "small world" properties where a small number of nodes or people have a very number of connection that can rapidly transmit contagion throughout the network [?]. It may be possible to apply network analysis to identify underlying structure of the contact network of opioid use and addiction, to identify pathways and points of contact between nodes or person in the spreading use, misuse, and abuse of prescription opioid medications. Future research could apply social network modeling to the opioid crisis in order to identify how drug dependancy and addiction are subserved by patterns of social interaction.

5 CONCLUSION

Several machine learning methods were used to classify heroin use and prescription opioid misuse and abuse. The results of this analysis are somewhat inconclusive, given that the direction of these effect is unknown. On the one hand there is evidence that individual who reported having used heroin were also more likely to report misusing or abusing prescription opioids. On the other hand, the proportion of individuals who misused or abused prescription opioids, and also reported using heroin, was twice as large as the proportion who reported only using heroin. A general conclusion is that the results provide partial support for the hypothesis that taking prescription opioids leads to a higher likelihood of illicid opioid use. However, the results did not provide sufficienbt evidence to rule out the alternative hypothesis that people who have used heroin may have a propensity for opioid use therefore be more likely to become dependent on prescription opioid medications. Given that the number of individuals who reported using heroin in this sample was low, additional data may help to provide evidence to resolve this question. A limitation of survey data is there may be bias in self-reports of illicit drug use, as it is a proscribed and illegal behavior, and therefore the data may underestimate the actual rate of heroin use in the general population. Including additional data from previous years may provide a more tests of these hypotheses.

Machine of opioid abuse can contribute to efforts to address prescription opioid addiction, overdoses, in the following ways:

(1) Identify factors related to opioid dependency
(2) Inform consumers of opioid medication as to risk factors
(3) Increase knowledge of opioid abuse for more informed prescriptions.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski, the Teaching Assistants, Juliette Zurick, Miao Jiang, Hungri Lee, Grace Li, Saber Sheybani Moghadam, and others who helped to improve this project and report.

REFERENCES

[1] Substance Abuse, Center for Behavioral Health Statistics Mental Health Services Administration, and Quality. 2016. *National Survey on Drug Use and Health (NSDUH) 2015.* Online data archive, United States Department of Health and Human Services., Ann Arbor, MI. https://doi.org/10.3886/ICPSR50011.v1

[2] Vittoria Colizza, Alain Barrat, Marc Barthlemy, and Alessandro Vespignani. 2006. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America* 103, 7 (2006), 2015–2020. https://doi.org/10.1073/pnas.0510525103 arXiv:http://www.pnas.org/content/103/7/2015.full.pdf

[3] Centers for Disease Control and Prevention. 2017. Prescription Opioid Overdose Data. online. (Oct. 2017). https://www.cdc.gov/drugoverdose/data/overdose.html

[4] hd1 and yoavram. 2016. Python: Download Returned Zip file from URL. Online. (Feb. 2016). https://stackoverflow.com/questions/9419162/python-download-returned-zip-file-from-url Stackoverflow.com.

[5] M. Herland, T. M. Khoshgoftaar, and R. Wald. 2014. A review of data mining using big data in health informatics. *Journal Of Big Data* 1, 2 (2014). https://doi.org/10.1186/2196-1115-1-2

[6] K. Johnson, A. Isham, D.V. Shah, and D.H. Gustafson. 2011. Potential Roles for New Communication Technologies in Treatment of Addiction. *Current psychiatry reports.* (2011). https://doi.org/10.1007/s11920-011-0218-y

[7] Rose A. Judd, Noah Aleshire, Jon E. Zibbell, and R. Matthew Gladden. 2016. *Increases in Drug and Opioid Overdose Deaths, United States, 2000-2014.* techreport 64(50). Centers for Disease Control and Prevention, Atlanta, GA. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6450a3.htm Morbidity and Mortality Weekly Report (MMWR).

[8] Lisa A. Marsch. 2012. Leveraging teachnology to enhance addiction treatment and recovery. *Journal of Addictive Diseases* 31, 3 (2012), 313–318. https://doi.org/10.1080/10550887.2012.694606

[9] Wes McKinney. 2017. *Python for Data Analysis.* O'Reilly Media Inc., Sebastopol, CA. https://github.com/wesm/pydata-book

[10] Andreas C. Muller and Sarah Guido. 2017. *Introduction to Machine Learning.* O'Reilly, Sebastopol, CA. https://github.com/amueller/introduction_to_ml_with_python/

[11] National Institute on Drug Abuse (NIDA). 2017. *Overdose Death Rates.* Summary. National Institutes of Health (NIH), Washington D.C. https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates

[12] Romualdo Pastor-Satorras and Alessandro Vespignani. 2001. Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett.* 86 (Apr 2001), 3200–3203. Issue 14. https://doi.org/10.1103/PhysRevLett.86.3200

[13] E. Rahm and H. Hai Do. 2000. *Data cleaning: Problems and current approaches.* techreport 23(4). Bulletin of the Technical Committee on Data Engineering, 1730 Massachusetts Avenue, Washington D.C. https://s3.amazonaws.com/academia.edu.documents/41858217/A00DEC-CD.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1511155930&Signature=VWRM7u4KWtP6ZxX5jB%2Bh6wMCbpg%3D&response-content-disposition=inline%3B%20filename%3DAutomatically_extracting_structure_from.pdf#page=5

[14] Sebastian Raschka and Vahid Mirjalili. 2017. *Ptyhon Machine Learning. Second Edition.* Packt, Birmingham, UK. https://github.com/rasbt/python-machine-learning-book-2nd-edition

[15] Yavin Shaham, Uri Shalev, Lin Lu, Harriet de Wit, and Jane Stewart. 2003. The reinstatement model of drug relapse: history, methodology and major findings. *Psychopharmacology* 168, 1 (01 Jul 2003), 3–20. https://doi.org/10.1007/s00213-002-1224-x

[16] S.M. Shiverick. 2017. BDA Project Data. github. (Nov. 2017). https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Project-Data.ipynb

[17] S.M. Shiverick. 2017. Classification Models of Heroin Use. github. (Nov. 2017). https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Analytics-Classifier-Heroin.ipynb Interactive Python Jupyter Notebook.

[18] S.M. Shiverick. 2017. Classification Models of Prescription Opioid Pain Relievers Misuse. github. (Nov. 2017). https://github.com/bigdata-i523/hid335/blob/master/project/BDA-Analytics-Classifier-PRL.ipynb Interactive Python Jupyter Notebook.

[19] S.M. Shiverick. 2017. Project Codebook for Data Variables from NSDUH-2015. github. (Nov. 2017). https://github.com/bigdata-i523/hid335/blob/master/project/project-data-codebook.txt

[20] J. Swendsen. 2016. Contributions of mobile technologies to addiction research. *Dialogues Clinical Neuroscience* 18, 2 (June 2016), 213–221. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4969708/

[21] Jake VanderPlas. 2017. *Python Data Science Handbook.* O'Reilly Media Inc., Sebastopol, CA. https://jakevdp.github.io/PythonDataScienceHandbook/

[22] Upkar Varshney. 2013. Smart medication management system and multiple interventions for medication adherence. *Decision Support Systems* 55, 5 (May 2013), 538–551. https://doi.org/10.1016/j.dss.2012.10.011

[23] Nora D. Volkow, Thomas R. Frieden, Pamela S. Hyde, and Stephen S. Cha. 2014. Medication-Assisted Therapies: Tackling the Opioid-Overdose Epidemic. *New England Journal of Medicine* 370, 22 (2014), 2063–2066. https://doi.org/10.1056/NEJMp1402780 arXiv:http://dx.doi.org/10.1056/NEJMp1402780 PMID: 24758595.

A CODE REFERENCES

All code, notebooks, files, and folders for this project can be found in the i523/hid335/project github repository: url:https://github.com/bigdata-i523/hid335/tree/master/project.

A.1 Download and Extract Data file

The 'get-data.py' function was written to download the data, unzip the data files, extract the data, and write the NSDUH-2015 dataset to CSV file [4].

A.2 Data Cleaning and Preparation

Data cleaning and preparation steps was conducted using an interactive python Jupyter Notebook [16] based on examples in Python for Data Analysis [9] and the Python Data Science Handbook [21].

A.3 Exploratory Data Analysis

Exploratory Data Analysis and Visualization was conducted using an interactive python notebook: CITE URL

based on examples from Python for Data Analysis [9], and the Python Data Science Handbook [21].

A.4 Machine Learning Classifier Algorithms

Machine learning classification models included logistic regression classifier, decision Tree classifier, random forests classifier, and gradient boosting classifier were constructed using SciKit Learn [10, 14] using two separate interactive python jupyter notebook, one for classifying Heroin Use as the target variable [17], and another notebook for classifying Prescription Opioid Misuse as the target [18].

List of Figures

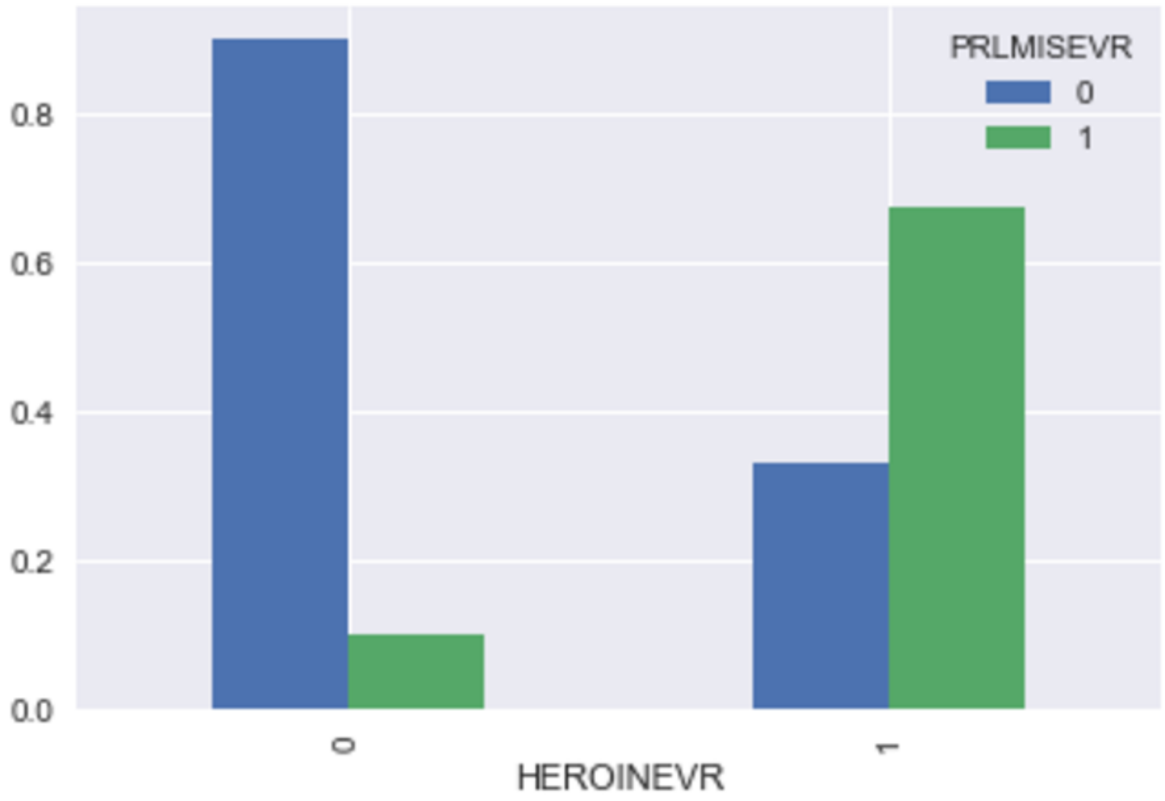

Figure 1: Proportion of Individuals Who Reported Ever Misusing Prescription Opioid Pain Relievers and Proportion Who Reported Using Heroin

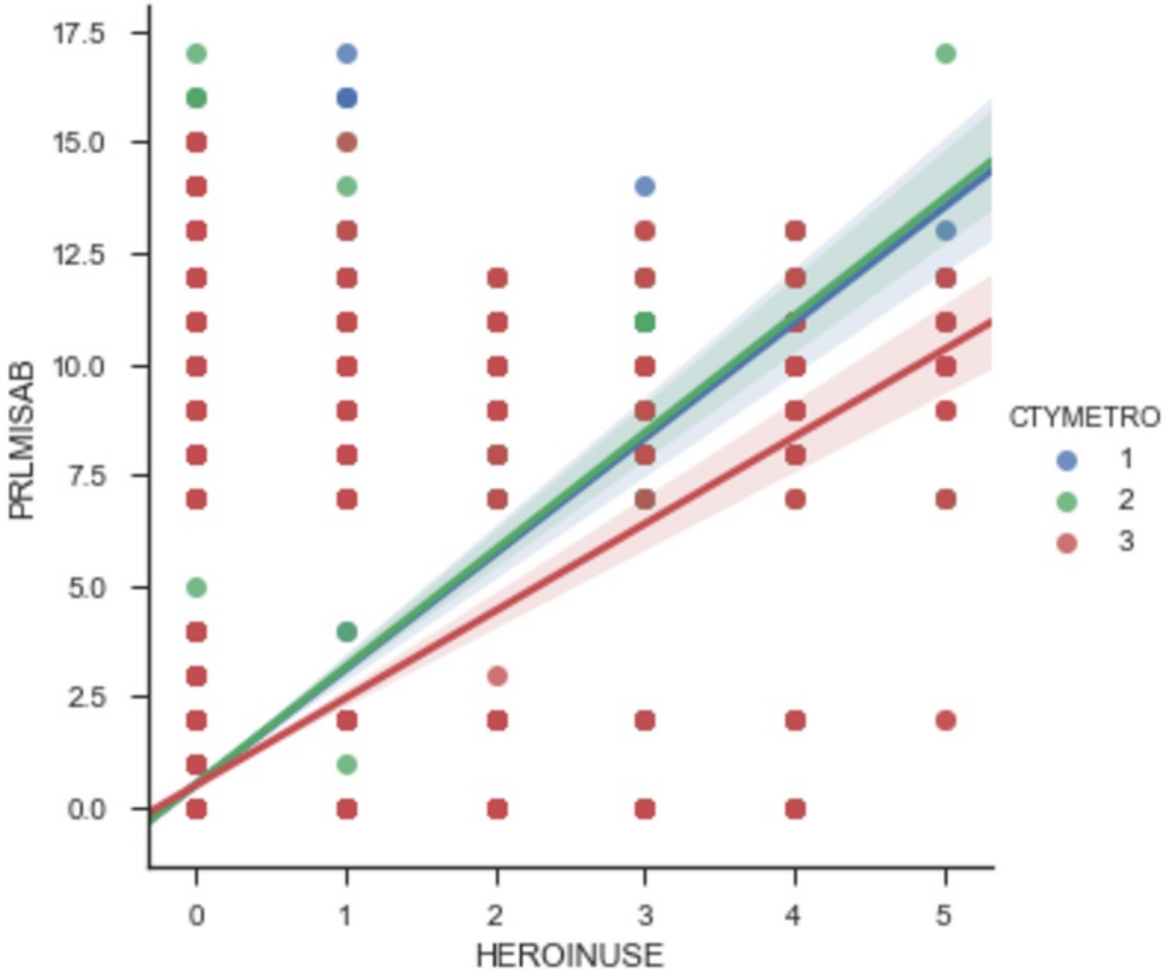

Figure 2: Plot of Opioid Pain Medication Misuse and Abuse and Heroin Use with Regression Slopes Weighted by Metropolitan Area Size

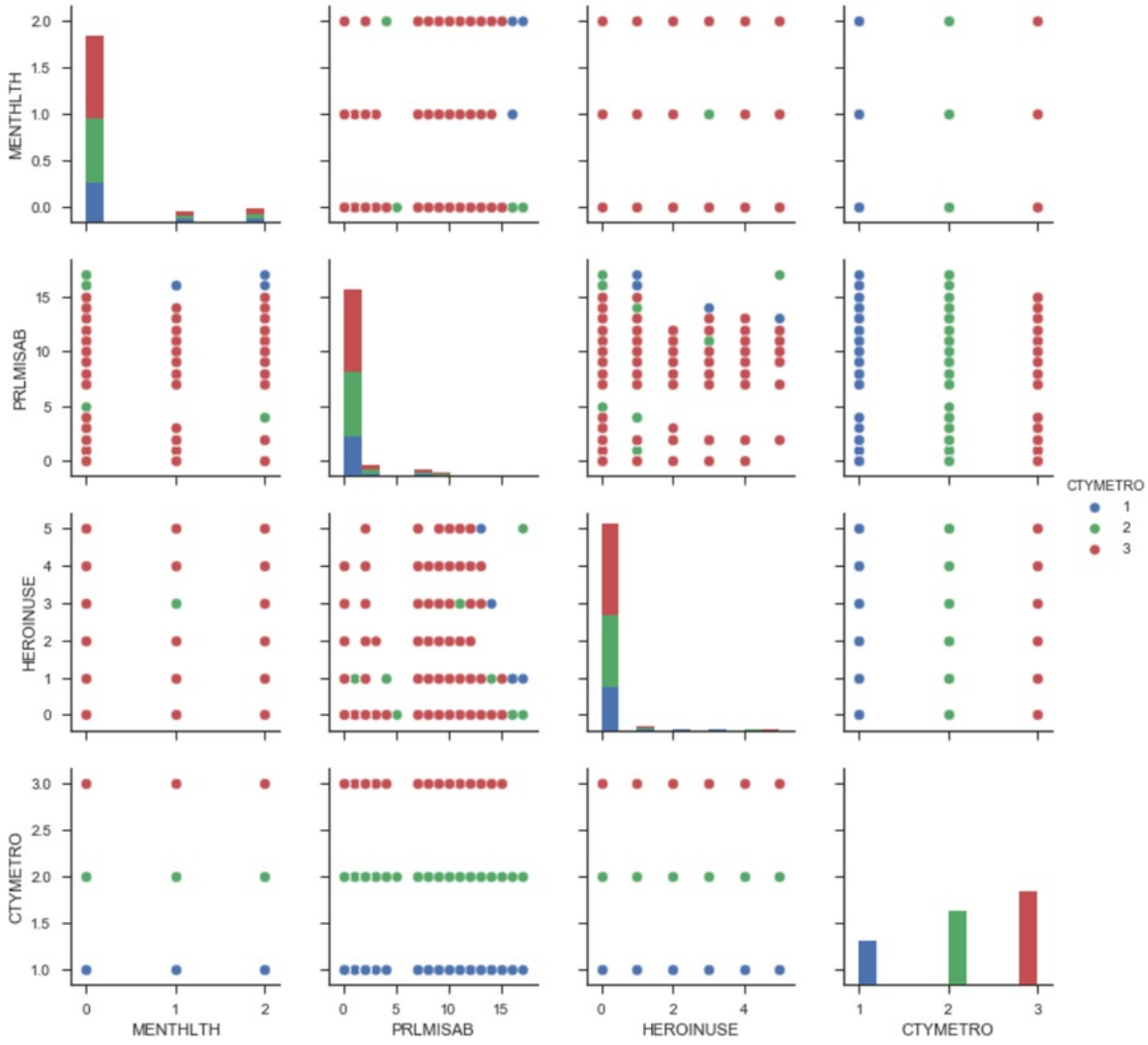

Figure 3: Pairplots of Mental Health, Prescription Opioid Misuse and Abuse, Heroin Use, and Size of City Metropolitan Area

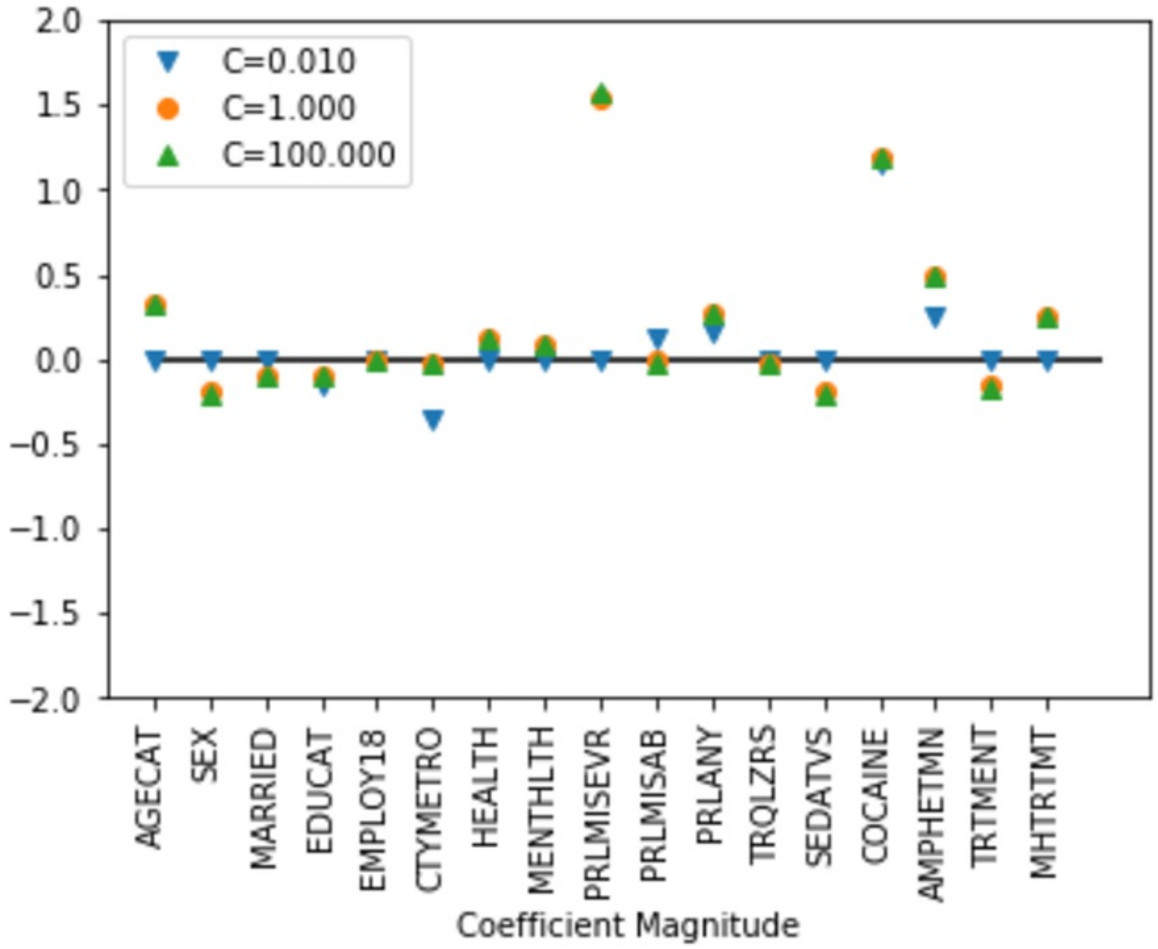

Figure 4: Coefficients of Logistic Regression Classifier of Heroin Use (With L1 Penalty and Values of Regularization Parameter C)

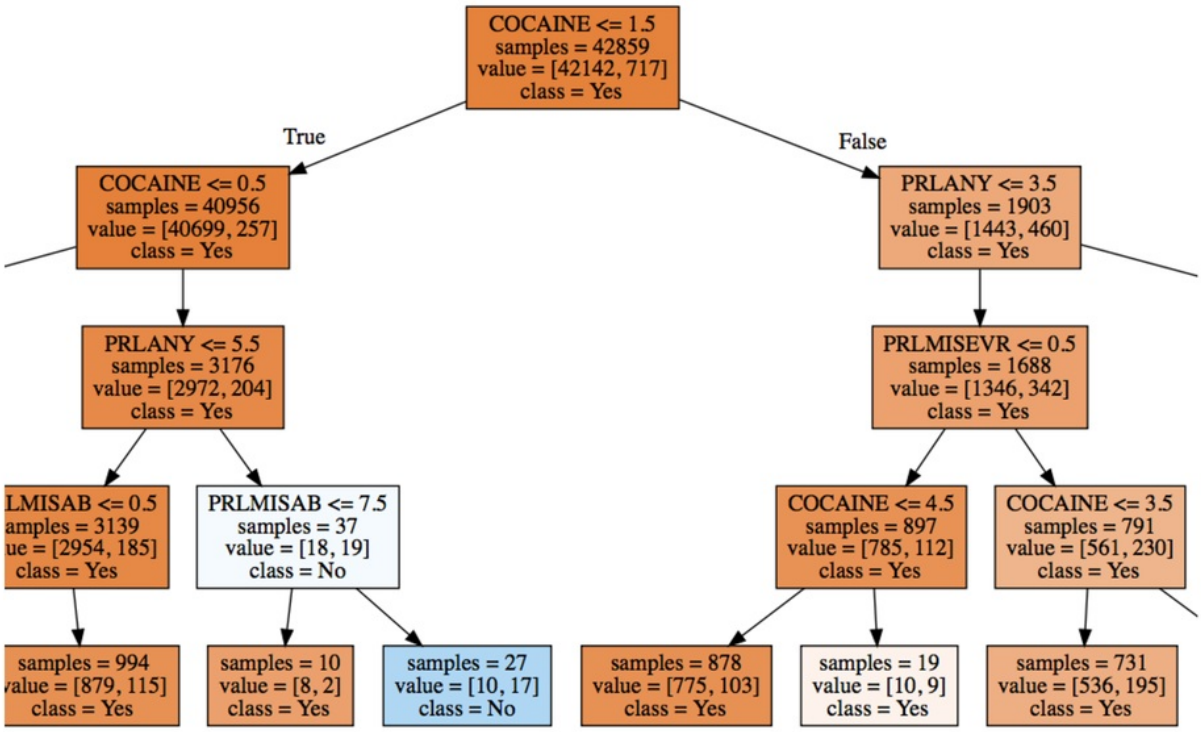

Figure 5: Decision Tree Classification of Heroin Use (Partial View)

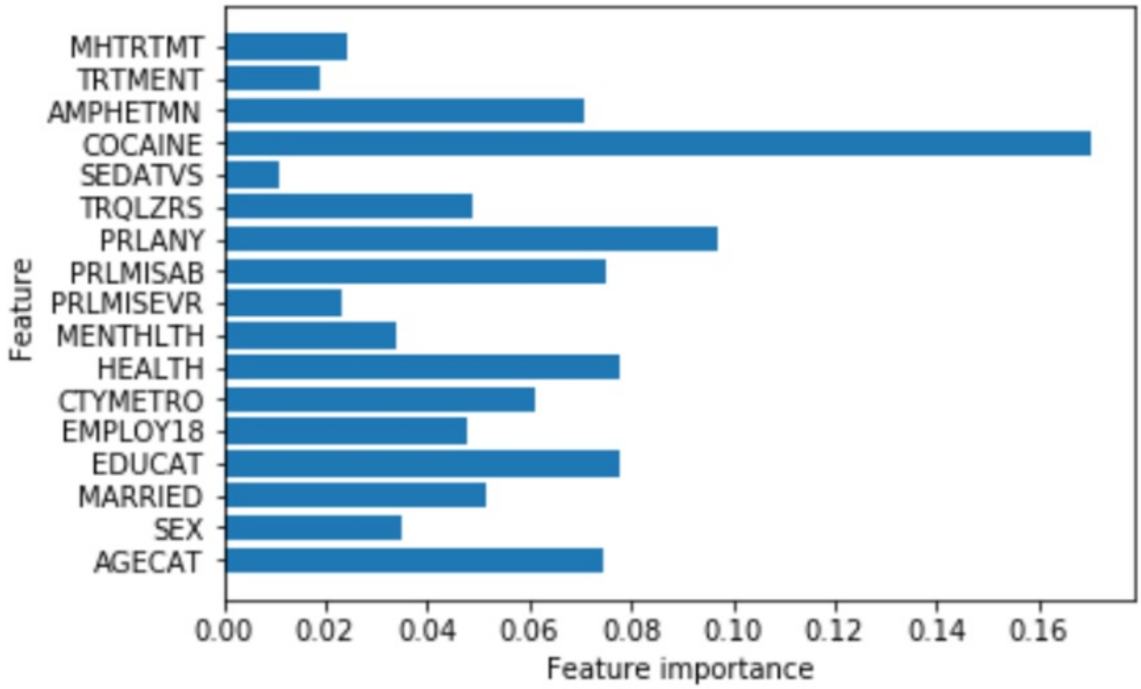

Figure 6: Feature Importance for Random Forests Classifier for Heroin Use

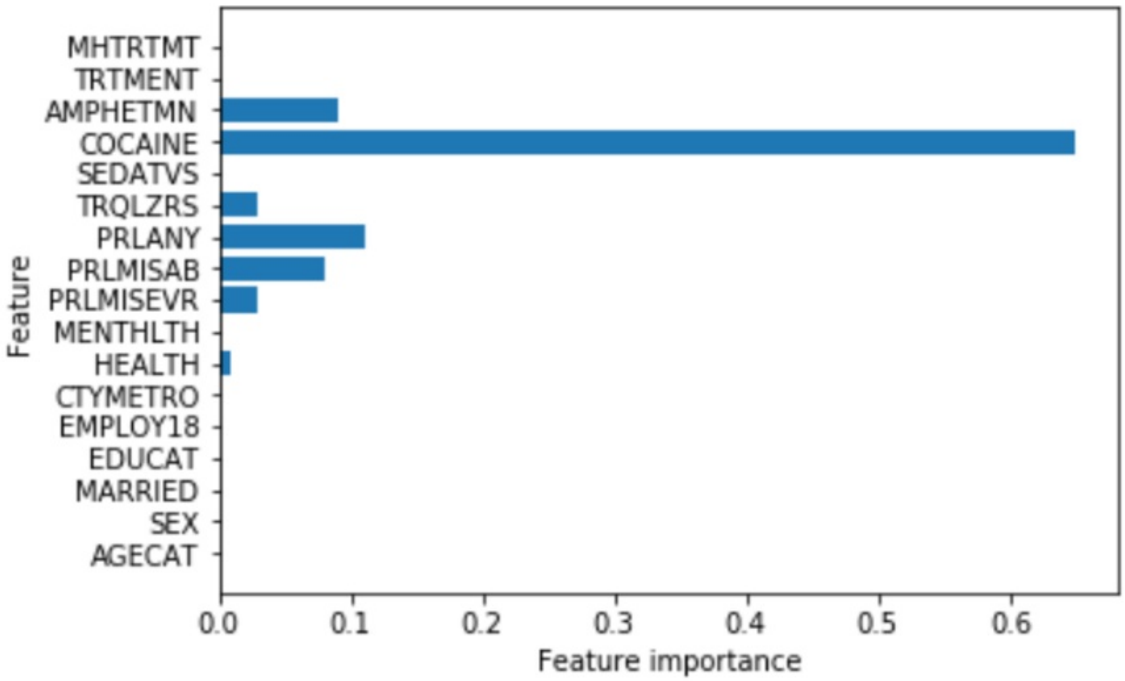

Figure 7: Feature Importance for Gradient Boosting Classifier for Heroin Use

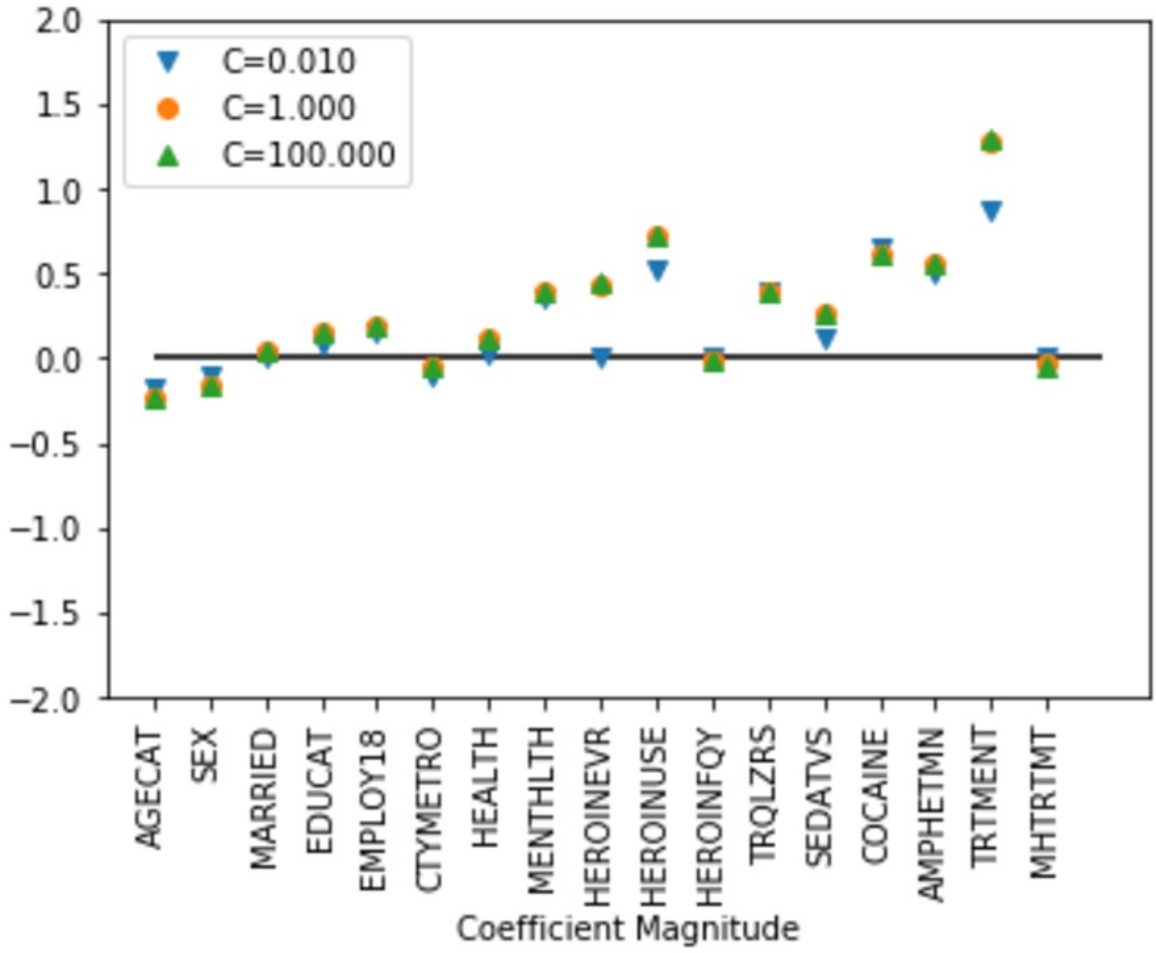

Figure 8: Logistic Regression Classification of Prescription Opioid (PRL) Misuse with L2 Penalty

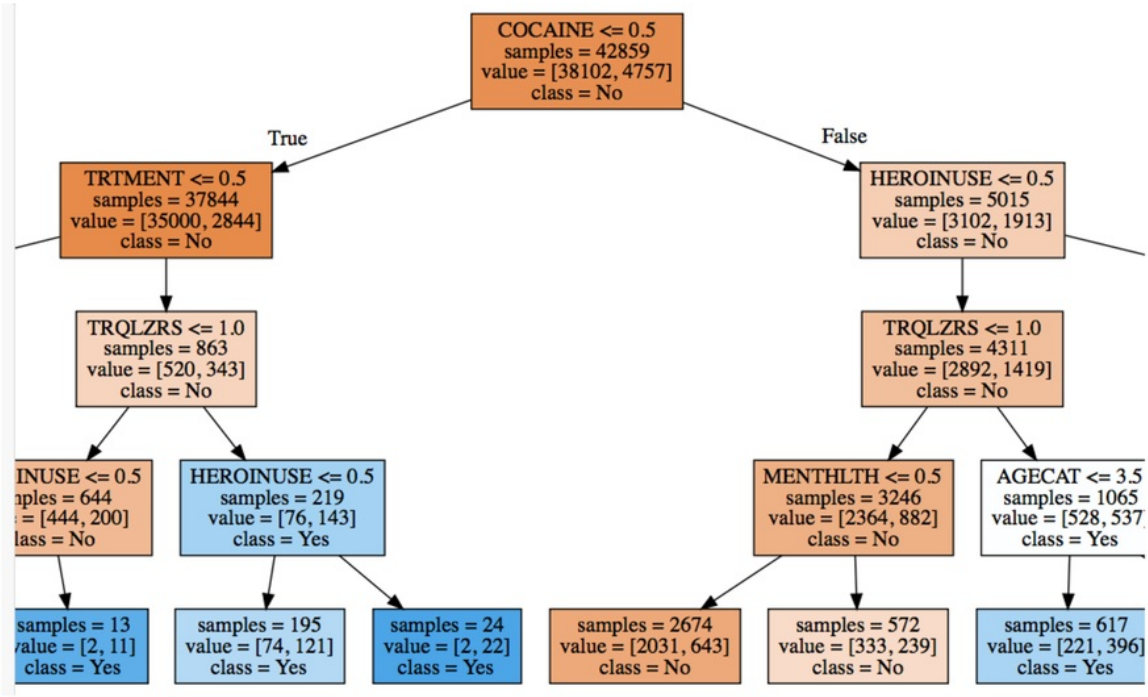

Figure 9: Decision Tree for Prescription Opioid (PRL) Misuse

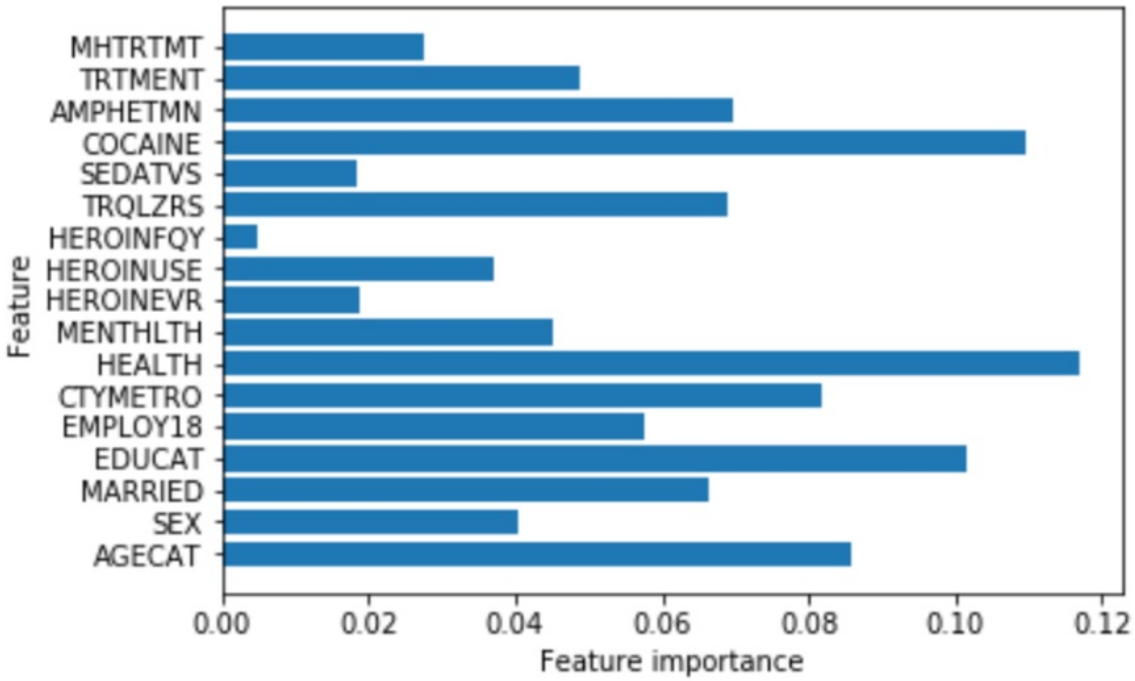

Figure 10: Feature Importance for Random Forest Classifier of Prescription Opioid (PRL) Misuse

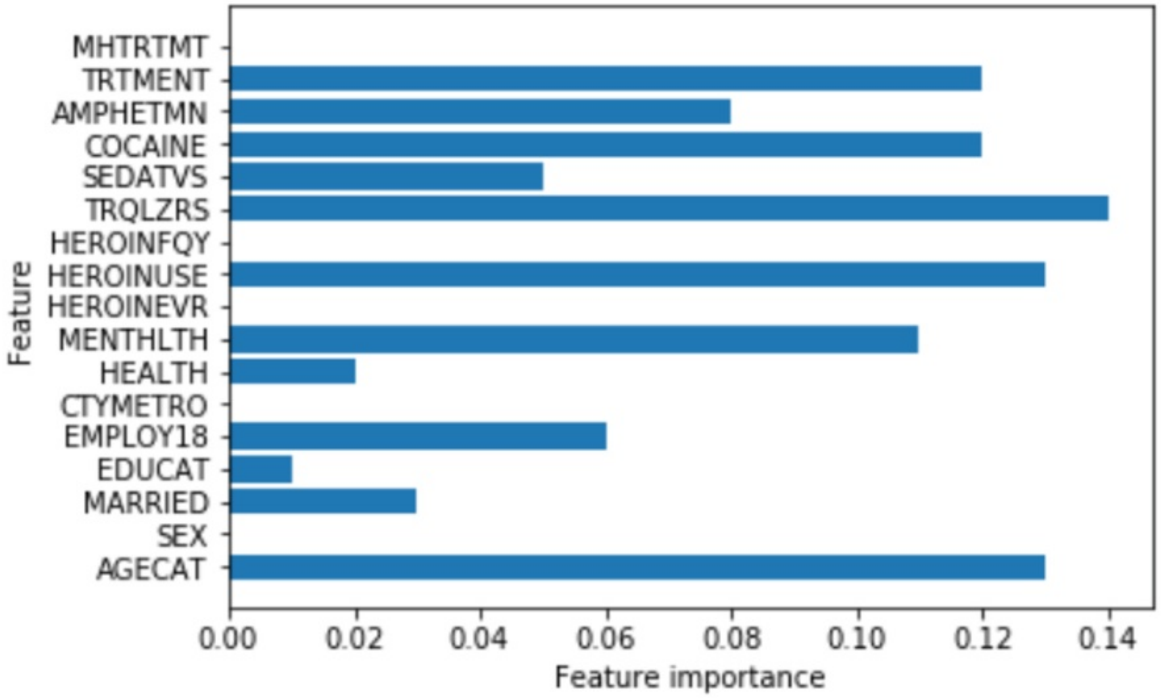

Figure 11: Feature Importance for Gradient Boosted Classifier Tree of Prescription Opioid (PRL) Misuse

List of Tables

Table 1: Substance Use by Age Group Counts - NSDUH 2015 [1]

Age Group	12-17	18-25	26-34	35-49	50+
Sample Size	13585	14553	9084	11169	8755
Oxycodone	545	1632	1132	1345	1044
Hydrocodone	831	2936	2233	2781	2103
Tramadol	241	753	654	829	734
Morphine	251	431	236	313	286
Fentanyl	28	97	81	96	86
Demerol	26	74	49	64	71
Buprenorphine	43	197	167	124	51
Oxymorphone	46	88	57	47	41
Hydromorphone	24	94	107	118	81
PRL Misuse Ever*	798	2127	1475	1343	600
Tranquilizers	405	1469	1064	1405	1153
Sedatives	204	242	157	256	226
Methadone Ever	32	83	96	71	46
Heroin Use Ever*	22	261	259	250	164
Cocaine Use Ever	109	1645	1626	1954	1406
Amphetamines Ever	932	1836	627	383	164
Methamphetamine	42	481	700	898	492
Hallucinogens	450	2660	2020	2127	1197
LSD Use Ever	190	1114	874	1442	907
Ecstasy (MDMA)	199	1867	1403	947	149

Table 2: Frequency Table of Mental Health Issues and Treatment NSDUH 2015 [1]

Age Group	12-17	18-25	26-34	35-49	50+
In Hospital Overnight	730	1149	821	890	1173
Adult Depression	0	2413	1395	1766	967
Suicidal Thoughts	13585	14553	9084	11189	8755
Mental Health Treatment					
Private Therapist	0	592	434	554	311
Treatment Gap*	469	931	321	239	90

bda_project

ORIGINALITY REPORT

9%	5%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Barbara Hammer. "Supervised Batch Neural Gas", Lecture Notes in Computer Science, 2006 Publication	1%
2	andre.panisson.com Internet Source	<1%
3	fas.org Internet Source	<1%
4	sociology-data.sju.edu Internet Source	<1%
5	Liaw, Chyn, Chun-Wei Tung, and Shinn-Ying Ho. "Prediction and Analysis of Antibody Amyloidogenesis from Sequences", PLoS ONE, 2013. Publication	<1%
6	Submitted to Associatie K.U.Leuven Student Paper	<1%
7	Submitted to Curtin University of Technology Student Paper	<1%

8	Submitted to Marian University Student Paper	<1%
9	Submitted to University of Wales, Bangor Student Paper	<1%
10	Dreyfus, C.. "A machine learning approach to the estimation of the liquidus temperature of glass-forming oxide blends", Journal of Non-Crystalline Solids, 20030401 Publication	<1%
11	files.eric.ed.gov Internet Source	<1%
12	"10-K: EGALET CORP.", EDGAR Online-Glimpse, March 11 2016 Issue Publication	<1%
13	aysps.wpdev.gsu.edu Internet Source	<1%
14	Submitted to Western Governors University Student Paper	<1%
15	Hossain, Syed Monowar, Amin Ahsan Ali, Md. Mahbubur Rahman, Emre Ertine David Epstein, Ashley Kennedy, Kenzie Preston, Annie Umbricht, Yixin Chen, and Santosh Kumar. "Identifying drug (cocaine) intake events from acute physiological response in the presence of free-living physical activity",	<1%

IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks, 2014.
Publication

16	Rudd, R. A., N. Aleshire, J. E. Zibbell, and R. Matthew Gladden. "Increases in Drug and Opioid Overdose Deaths-United States, 2000-2014", American Journal of Transplantation, 2016. Publication	<1%
17	Margarita Villagrana, Sei-Young Lee. "Parental Influence on Adolescent Cigarette, Alcohol and Marijuana Use: A Focus on Race/Ethnicity and Age", Child and Adolescent Social Work Journal, 2017 Publication	<1%
18	Submitted to The University of Manchester Student Paper	<1%
19	Advances in Intelligent Systems and Computing, 2016. Publication	<1%
20	Submitted to University College London Student Paper	<1%
21	dblp.uni-trier.de Internet Source	<1%
22	Submitted to University of Utah	

	Student Paper	<1%
23	Submitted to Widener University Student Paper	<1%
24	www.oas.samhsa.gov Internet Source	<1%
25	Submitted to Brigham Young University Student Paper	<1%
26	Rudd, Rose A.; Aleshire, Noah; Zibbell, Jon E. and Gladden, R. Matthew. "Increases in Drug and Opioid Overdose Deaths -- United States, 2000-2014", MMWR: Morbidity & Mortality Weekly Report, 2016. Publication	<1%
27	drugabusestatistics.samhsa.gov Internet Source	<1%
28	www.cdc.gov Internet Source	<1%
29	researcharchive.vuw.ac.nz Internet Source	<1%
30	arxiv.org Internet Source	<1%
31	d-scholarship.pitt.edu Internet Source	<1%

32	blog.caradigm.com Internet Source	<1%
33	www.scirp.org Internet Source	<1%
34	somatosphere.net Internet Source	<1%
35	Francisco Herrerías-Azcué, Tobias Galla. "The effects of heterogeneity on stochastic cycles in epidemics", Scientific Reports, 2017 Publication	<1%
36	addiction.surgeongeneral.gov Internet Source	<1%
37	dsplab.eng.fiu.edu Internet Source	<1%
38	uhdspace.uhasselt.be Internet Source	<1%
39	digitalcommons.lmu.edu Internet Source	<1%
40	journals.aps.org Internet Source	<1%
41	pesquisa.bvsalud.org Internet Source	<1%
42	orbilu.uni.lu Internet Source	<1%

43	Trudeau, Kimberlee J.. "Development of a community readiness survey for coalitions to address prescription opioid misuse.(Su", Journal of Alcohol & Drug Education, Dec 2015 Issue Publication	<1%
44	Bailey, J.E.. "The effect of FDA approval of a generic competitor to OxyContin^(R) (oxycodone HCl controlled-release) tablets on the abuse of oxycodone", Drug and Alcohol Dependence, 20060915 Publication	<1%
45	www.cbr.washington.edu Internet Source	<1%
46	Albert, Mark V., Konrad Kording, Megan Herrmann, and Arun Jayaraman. "Fall Classification by Machine Learning Using Mobile Phones", PLoS ONE, 2012. Publication	<1%
47	Janani, Hamed, Nathan D. Jacob, and Behzad Kordi. "Automated recognition of partial discharge in oil-immersed insulation", 2015 IEEE Electrical Insulation Conference (EIC), 2015. Publication	<1%
48	Sunghee Lee, Tuba Suzer-Gurtekin, James Wagner, Richard Valliant. "Total Survey Error	<1%

and Respondent Driven Sampling: Focus on Nonresponse and Measurement Errors in the Recruitment Process and the Network Size Reports and Implications for Inferences", Journal of Official Statistics, 2017

Publication

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		