

Comparing Predictive Models of Pain Reliever Misuse and Abuse

Sean M. Shiverick
Indiana University-Bloomington

ABSTRACT

The misuse and abuse of prescription opioids (MUPO) has become a major health crisis in the U.S. Predictive modeling provides a useful approach for analyzing pain reliever misuse and abuse and identifying features that contribute to MUPO. This study compared ten classification models using four performance metrics: accuracy, sensitivity (i.e. recall), precision, and f_1 -score. Data from three years of the National Survey on Drug Use and Health (2015-2017) was combined into a sample of $N = 170317$ observations. Twenty-six percent of respondents reported using prescription opioid medications in the past year, 11% reported misusing or abusing pain relievers, and 2% reported heroin use. The classifier models were fit to a training set using sixteen features that included demographic variables, medications, and illicit drugs. The binary target variable was pain reliever misuse and abuse. Model performance was evaluated on the testing set. The f_1 -score was used as the performance metric due to unbalanced classes in the data. Logistic regression, random forests, and decision trees had the highest f_1 -scores compared to other models. All three models identified cocaine use as the most informative feature for predicting pain reliever misuse and abuse. Amphetamine use was selected as the second most important variable by logistic regression and random forests. The models differed in the importance they assigned to heroin use, tranquilizer use, age category, and mental health as predictors of pain reliever misuse and abuse. Advantages and limitations of the classifier models are considered and the tradeoff between model complexity and interpretability is discussed.¹

KEYWORDS

Predictive Modeling, Supervised Learning, Classification Models

1 INTRODUCTION

Over the past two decades the misuse and abuse of prescription opioids (MUPO) has become a major health crisis in the U.S. [1]. In 2015, an estimated 2 million Americans suffered a substance use disorder related to prescription opioid pain relievers such as oxycodone or hydrocodone [2]. Opioid dependence and addiction are chronic health conditions. On average, more than 115 people die from an opioid overdose in the U.S. each day [3], and opioid-related overdose deaths have more than quadrupled from 1999 to 2016. In addition, non-medical use of prescription opioids is a significant risk factor for heroin use [4]. Supply-based interventions to reduce the availability of prescription opioids have produced a shift to the use of heroin and synthetic opioids such as fentanyl [5]. Four in five new heroin users started out misusing prescription painkillers [6]. Given that the potency and dosage levels of illicit or synthetic opioids is largely unknown, the risk of overdose death is greatly

increased. The sharp rise in prescription overdose deaths (POD) and heroin overdose deaths (HOD) are correlated [7, 8]. Predictive modeling provides useful methods for analyzing the misuse of prescription opioids and identifying features that contribute to MUPO. Data mining can predict individuals who may be susceptible to opioid addiction and provide insights to inform policy decisions for addressing the opioid crisis. This study compares the performance of several classification models to determine the best approaches for modeling pain reliever misuse and abuse.

1.1 Predictive Modeling

Predictive modeling, statistical learning, or machine learning describe a set of procedures and automated processes for extracting knowledge from data [9–12]. The two main branches of predictive modeling are supervised learning and unsupervised learning. Supervised learning problems involve prediction about a specific target or outcome variable. If a dataset has no target outcome, unsupervised learning methods can help to reveal underlying structure in the data (e.g. clustering). Supervised learning is used to predict an outcome based on input provided to a model, when examples of input/output pairs are available in the data [11]. A statistical learning model is constructed using a set of observations to train the model and then make predictions with new observations. Two main approaches for supervised learning problems are regression and classification. When the target variable is continuous or there is continuity in the outcome (e.g. home prices), a regression model tests how a set of features predicts the target variable. If the target is a class label, binary variable, or set of categories (e.g., spam or ham emails, benign or malignant cells), a classification model will predict which class or category label new instances are assigned to. This study used a supervised learning approach to classify prescription pain reliever misuse and abuse as a binary outcome.

In the era of ‘big data’, large amounts of health information are being generated from electronic medical records (EMRs), clinical research data, and population-level health data [13]. Although it can be difficult to obtain reliable information about opioid use based on self-reports, surveys provide data on a range of issues that people may be reluctant to disclose such as illicit drug use and mental health problems. The data for the present study was obtained from the National Survey on Drug Use and Health (NSDUH) which is a major source of information for the use of illicit drugs and mental health issues among the U.S. population aged 12 or older [14]. The NSDUH is a comprehensive public survey that includes a diverse array of questions and variables related to the use, misuse, and abuse of substances including alcohol, tobacco, prescription medications, and illicit drugs. In addition to typical demographic information, the survey includes self-reported measures on items related to physical health, mental health (e.g., depression, anxiety, suicide), as well as counseling, and drug or alcohol treatment. Data from the NSDUH has been used for identifying groups at high risk for substance use and the co-occurrence of substance use and mental health disorders.

¹This project was completed in partial fulfillment of requirements for the M.S. in Data Science from the School of Informatics and Computing at IU-Bloomington completed in May, 2018. The manuscript was completed November 23, 2018 and submitted December 15, 2018. Address correspondence to smshiver@iu.edu.

The target variable for the study was any previous misuse or abuse of prescription opioid pain relievers (e.g., oxycodone, hydrocodone). The predictor variables were demographic features (e.g., age, sex, education, etc.), use of medications (e.g., tranquilizers, sedatives), and illicit drugs (e.g., cocaine, amphetamines, heroin).

1.2 Classification Models

1.2.1 Linear Models. As stated by the statistician George E. P. Box, "All models are wrong, but some models are useful" [15]. There are advantages and limitations to selecting any model. Logistic regression is one of the most reliable and interpretable models for classification. Logistic regression models the conditional distribution of probabilities for a binary response (e.g., $Pr(Y = k|X = x)$) as a combination of a set of predictor variables [9, 12]. The decision boundary for the logistic regression classifier is a linear function of the input; a binary classifier separates two classes using a line, plane, or hyperplane [11]. Given that the probability values for the outcome range between 0 and 1, predictions can be made based on a default value. For example, a default value of 'Yes' could be predicted for any individual for whom the probability of pain reliever misuse and abuse is greater than fifty-percent: $Pr(PRLMISAB) > 0.5$. Logistic regression uses a maximum likelihood method to predict the coefficient estimates that correspond as closely as possible to the default state. In other words, the model will predict a number close to one for individuals who have misused or abused pain relievers and a number close to zero for individuals who have not. The distribution of conditional probabilities in the logit model has an S-shaped curve. By taking the natural log of the left side of the regression equation, we obtain the logit function which is linear in the parameters, where $X = (X_1 \dots X_p)$ predictor variables (equation 1). The coefficient estimates are selected to maximize the likelihood function, and are interpreted as an indication of the log-odds change in the outcome variable that is associated with a one-unit increase in the predictor variable, holding the effects of other predictor variables constant. The intercept, β_0 , is the log of the odds ratio when X is 0 [16].

$$\ln \frac{P_i}{1 - P_i} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (1)$$

Linear discriminant analysis (LDA) is an alternative approach to estimating probabilities that models the distribution of predictors separately in each of the response classes and then uses Bayes' theorem to 'flip' these into estimates for, $Pr(Y = k|X = x)$ [9]. The term linear in LDA refers to the discriminant functions being linear functions of the predictors. For distributions assumed to be normal (i.e., multivariate Gaussian), LDA provides a model that is similar in form to logistic regression, but more stable. LDA is also preferred for outcomes with more than two response classes. The important assumptions for LDA are, first, a common covariance matrix for all classes, and second, the class boundaries are linear functions of the predictors. *Quadratic Discriminant Analysis* (QDA) is an approach that assumes each class has its own covariance matrix and the decision boundaries are quadratically curvilinear in the predictor space [10]. LDA is less flexible as a classifier than QDA, but can perform better with relatively few training observations or when the majority of predictors in the data represent discrete categories. QDA is recommended over LDA with a very large

training set or when the decision boundary between two classes is non-linear.

1.2.2 Non-linear Models. The performance of linear classifiers suffers when there is a non-linear relationship between the predictors and target outcome. With training observations that can be separated by hyperplane, the maximal marginal classifier provides the maximum distance (i.e., margin) from each observation to the hyperplane [9]. The test observations are classified based on which side of the hyperplane they fall; however, in many cases no separating hyperplane exists. The *support vector classifier* (SVC) extends the maximal margin classifier by using a soft margin that allows a small number of observations to be misclassified on the wrong side of the hyperplane [10, 17]. The observations that fall directly on the margin or on the wrong side of the hyperplane are called 'support vectors'. The parameter 'C' indicates the number of observations that can violate the margin; if $C > 0$, no more than C observations can be on the wrong side of the hyperplane. SVC addresses the problem of non-linear boundaries between classes by enlarging the feature space with higher order (e.g., quadratic, cubic, polynomial) functions of the predictors.

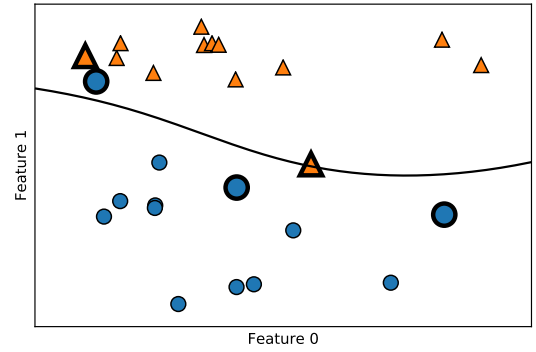


Figure 1: Decision Boundary and Support Vectors for Support Vector Machine (SVM) with Nonlinear Kernel [11]

Support vector machines (SVM) are an extension of SVC that use a kernel trick to reduce computational load. The radial basis function (RBF) kernel (i.e., Gaussian kernel) is one of the most commonly used approaches. In training the model, only a subset of data points is used to construct the decision boundary, namely the support vectors that lie on the border that separates the two classes. In predicting classes for new observations, the algorithm calculates the distance to each of the support vectors measured by the Gaussian kernel [11]. Figure 1 shows an example of the non-linear decision boundary obtained with SVM using the RBF kernel; the decision boundary is a smooth curve and the support vectors are the large points in bold outline. Even with the default settings, the RBF kernel provides a decision boundary that is decidedly non-linear. The parameters for SVM are 'C', which regulates the importance of each data point, and 'gamma' which controls the width of the Gaussian kernel. A small value of C indicates a restricted model in which the influence of each data point is limited and the algorithm

adjusts to the majority of data points. With larger values of C , more importance is given to each data point and the model tries to correctly classify as many training observations as possible, which results in more curvature in the decision boundary. Large values of γ mean that only close values are relevant for classification, resulting in a smooth decision boundary. Small values of γ mean that far points are similar. If the values of both C and γ are large, each point can have a large influence in a small region, which produces a choppy decision boundary. If the values of C and γ are both small, the decision boundary becomes close to linear.

Bayes' rule was mentioned above in relation to LDA; in this section, the *Naive Bayes Classifier* is considered as a non-linear model. The Bayes theorem (equation 2) is represented by a set of probabilities that answer the question, "Based on a given set of predictors, what is the probability that an outcome belongs to a particular class?"

$$P(Y = cl|X) = \frac{P(Y) * P(X|Y = cl)}{P(X)} \quad (2)$$

The prior probability, $P(Y)$, is the expected probability of a given class based on what is known (e.g. rate of disease in the population). $P(X)$ is the probability of the predictor variables. The conditional probability, $P(X = cl|Y)$, is the probability of observing the predictor variables for data associated with a given class. The naive Bayes model assumes that all of the predictor variables are independent, which is not always realistic. The conditional probabilities are calculated based on the probability densities for each individual predictor [10]. For categorical predictors, the observed frequencies in the training set data can be used to determine the probability distributions. The prior probability allows us to tilt the final probability toward a particular class. Class probabilities are created and the predicted class is the one that is associated with the largest class probability. Despite the somewhat unrealistic assumption of independence among predictors, the naive Bayes model is computationally quick, even with large training sets, and performs competitively compared to other models. The naive Bayes model encounters issues when dealing with frequencies or probabilities equal to zero, especially for small sample sizes. In addition, as the number of predictors increases relative to the size of the sample, the posterior probabilities will become more extreme.

Neural Networks are powerful models for classification and regression that are based on theories of connectivity in the brain [10]. The present study considers a simple method called a multilayer perceptron (MLP) which is a feed-forward neural network [11, 12]. The outcome is modeled by an intermediary set of unobserved variables called hidden units, which are linear combinations of the original predictors (see Appendix). Each hidden unit is a combination of some or all of the predictors which are then transformed by a nonlinear function (e.g. sigmoidal). A neural network usually has multiple hidden units used to model the outcome. The MLP classifier computes weights between the inputs and hidden units, and weights between the layers of hidden units and the output. After computing each hidden unit, the output is modeled by a nonlinear combination of the hidden units. The nonlinear function allows the neural network to fit more complicated functions than a linear model; however, neural networks are sensitive to the scaling of

the features and can require extensive data preprocessing. There are several ways to modify the complexity of a neural network: by selecting the number of hidden layers, the number of units within each layer, and the regularization parameter (L2) which shrinks the weights towards zero. The feature weights provide an estimate of feature importance. Although neural networks can capture information in large amounts of data with very complex models, they tend to overfit data used to train the model and can be difficult to interpret. Neural networks may work best with homogenous datasets where the predictor variables all have similar meanings [7]. For datasets with many different kinds of features, tree-based methods offer a better approach.

1.2.3 Tree-based Models. Decision trees are based on a hierarchy of 'if-else' questions starting from a root node and proceeding through a series of binary decisions or choices. Each node in the tree 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 continuous features, questions are expressed in the form: "Is feature x larger than value y ?" In constructing the tree, the algorithm searches through all possible tests and finds a solution that is most informative about the target outcome [11]. The recursive branching process yields a binary tree of decisions, with each node representing a test for a single feature. This process of 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 models is that they require little adjustment and are easy to interpret. A drawback is that they can lead to complex models which are highly overfit to the training data. 'Prepruning' can help reduce overfitting by limiting the maximum depth of the tree, or the maximum number of leaves. Another approach is to set the minimum number of points in a node required for splitting. Decision trees work well with features measured on different scales, or with data that has a mix of binary and continuous features.

Random Forests is an ensemble approach that combines many simple trees that each overfit the data in different ways. By building many trees and averaging their results, random forests help to reduce overfitting. In constructing the forests, the user selects the number of trees to build (e.g., 1000). Randomness is introduced using a bootstrapping method that repeatedly draws random samples of size n from the data set, with replacement. The decision trees are built on these random samples of the same size, with some points missing and some data points repeated [11, 12]. The algorithm makes a random selection of p -features, and uses a different set of features at each node branch. These processes ensure that all of the decision trees in the random forest are different. Random forests is one of the most widely used supervised learning algorithms and works well without very much parameter tuning or scaling of data. A limitation is that random forests do not perform well with high-dimensional data, or data that is sparse (e.g., text).

Gradient Boosting is another ensemble method that combines multiple decision trees in a serial fashion, where each tree tries to correct for mistakes of the previous one [11]. Gradient boosted regression trees use strong prepruning, with shallow trees of a

Table 1: Confusion Matrix for Binary Classification.

Predicted	Actual Outcome	
	Outcome	
No Misuse	No Misuse	True Negative (TN)
	PRL Misuse	False Negative (FN)
PRL Misuse	No Misuse	False Positive (FP)
	PRL Misuse	True Positive (TP)

depth of one to five. Each tree only provides a good estimate of part of the data; combining many shallow trees (i.e., “weak learners”) 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. A common approach is to first perform random forests and then include gradient boosting to improve model accuracy.

1.3 Evaluating Model Performance

To identify which model is best for a given problem, with the data available, it is necessary to evaluate the performance of different learning algorithms. Binary classification is assessed in terms of the successful assignment of observations to one of two classes: positive or negative. No classification model can make perfect predictions, as errors are always to be found. Medical testing is often used as an example to illustrate classification decisions and errors. For example, in the actual state of the world, a patient either has an illness or not, and the person is either diagnosed as having the illness or not. In the present case, the positive class represents self-reported pain reliever misuse and abuse (PRL Misuse), and predictions based on the classifier models will be either correct or incorrect in relation to the observed outcomes. For example, a person who has never misused or abused pain relievers may be misclassified as having done so (i.e., ‘false positive’), or conversely, a person who actually has misused and abused pain relievers may be mislabeled as never having done so (i.e., ‘false negative’).

Correct decisions and classification errors are represented in a *confusion matrix* that indicates the correspondence between predicted and actual outcomes (Table 1). The confusion matrix is a two-by-two array in which the columns correspond to the actual observed classes and the rows correspond to the predicted classes. The main diagonal indicates the number of correctly classified samples (i.e., true negative, true positive), while the other entries represent the number of samples in one class that were mistakenly classified as another class. Several performance metrics can be obtained from the confusion matrix, including accuracy, sensitivity (i.e., recall), specificity, precision, and the f_1 -score, presented in Table 2 [10, 18]. Model performance is most commonly evaluated using *Accuracy* which is assessed by the number of correct predictions divided by the total observations. Sensitivity or *Recall* provides the “True Positive Rate” (TPR), measured as the number of positive samples that

Table 2: Performance Metrics for Classifier Models.

Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
Sensitivity, Recall (TPR)	$\frac{TP}{TP+FN}$
Specificity (TNR)	$\frac{TN}{TN+FP}$
Precision (PPV)	$\frac{TP}{TP+FP}$
f_1 score	$2 * \frac{Precision * Recall}{Precision + Recall}$

are correctly identified by the prediction (number of sick patients correctly diagnosed). Recall is used when the goal is to avoid false negatives. The True Negative Rate (TNR) is described as Specificity, which indicates the proportion of negative cases correctly identified (healthy people not misdiagnosed). *Precision* provides the “Positive Predictive Value” (PPV), which measures how many of the samples predicted as positive are actually positive. Precision is used as a metric when the goal is to limit the number of false positives. The f_1 -score represents a harmonic mean between recall and precision ($\frac{2}{\frac{1}{R} + \frac{1}{P}}$). The f_1 -score can be a better measure of performance than accuracy in datasets with imbalanced classes, where one class is much more frequent than the other class, as it takes into account both recall and precision [11, 19].

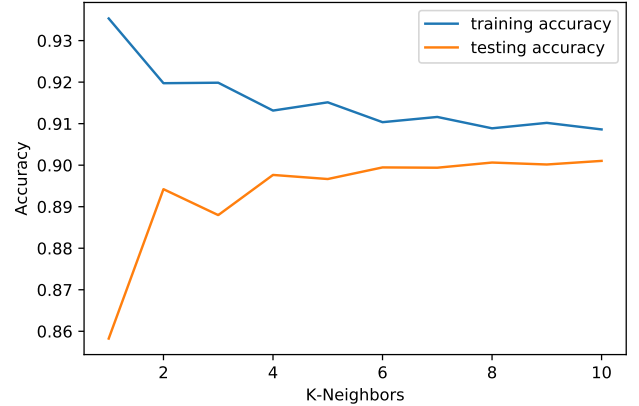


Figure 2: Training Set Accuracy and Testing Set Accuracy for KNN Classifier Model as a function of K-Neighbors.

1.3.1 Training Accuracy and Test Set Accuracy. In constructing and evaluating predictive models it is important to select a model that performs well not only with the data used to train the model, but also with new observations. A standard practice is to divide the sample data into a *training set* and a *testing set* of observations that is set aside and used to evaluate model performance. By convention, approximately 70 to 80 percent of observations are used in the training set and the remaining portion is held in the test set. Two main problems in evaluating model performance are *overfitting*

Table 3: Variables Included in the Sample Data for Model Construction.

Target Variable	Label
Prescription opioid pain reliever misuse and abuse (Likert scale: 0-12)	PRLMISAB
<i>Predictor Variables</i>	
Year of NSDUH survey (15=2015, 16=2016, 17=2017)	YEAR
Age category (1=12-17 years, 2=18-25, 3=26-35, 4=36-49, 5=50 and older)	AGECAT
Sex (0=Male, 1=Female)	SEX
Marital status (0=unmarried, 1=divorced, 2=widowed, 3=married)	MARRIED
Education level (1=h.s. or Less, 2=h.s. grad., 3=some college, 4=college grad.)	EDUCAT
Employment status, over age 18 (1=not employed, 2=part-time, 3=full-time)	EMPLOY18
Size of city/metropolitan region (1=rural, 2=small, 3=large)	CTYMETRO
Health problems, aggregated (Likert scale: 0-10)	HEALTH
Mental health, aggregated: adult depression, emotional distress (Likert scale: 0-10)	MENTHLTH
Treatment for drugs or alcohol in past year, aggregated (Likert scale: 0-5)	TRTMNT
Mental health treatment, aggregated (Likert scale: 1-10)	MHTRTMT
Tranquilizer use in past year, aggregated (Likert scale: 0-5)	TRQLZRS
Sedative use in past year, aggregated (Likert scale: 0-5)	SEDATVS
Heroin use in past year, aggregated (Likert scale: 0-5)	HEROINUSE
Cocaine and crack cocaine use in past year, aggregated (Likert scale: 0-5)	COCAINE
Amphetamine and methamphetamine use in past year, aggregated (Likert scale: 0-5)	AMPHETMN

and *underfitting*. In the case of overfitting, a model can have high accuracy on the training set but perform poorly with new data in the test set because the model is overly fit to the training data. By contrast, in the case of underfitting, a simple model may not generalize well to new observations as it does not include all of the features relevant for predicting the target outcome. One of the simplest classification models, K-Nearest Neighbors (KNN) provides an example of the tradeoff between training accuracy and test set accuracy. KNN classifies observations by assigning the label that is most frequent among the ' k '-number of nearest training samples (k is a parameter selected by the user). The accuracy of the KNN classifier for the training set and testing set is plotted in Figure 2 as a function of the parameter k -neighbors. The plot shows that increased accuracy on the training set is associated with lower testing set accuracy; conversely, increased accuracy on the testing set is related to a decrease in training set accuracy. The ideal model is one that optimizes test set accuracy while striking a balance between the problems of overfitting and underfitting. In the case of KNN, testing set increases slightly between 2 and 4 neighbors, but does not improve much beyond 5 neighbors. Therefore, a model with $k=4$ neighbors provides a reasonable solution for the data.

1.3.2 Imbalanced Classes. Previous studies have analyzed the misuse and abuse of prescription opioids (MUPO) using logistic regression and identified factors that influence MUPO such as gender and mental illness [5, 8, 20, 21]. The present study extends previous work by comparing the performance of ten classifier models of pain reliever misuse and abuse and evaluating each model using four performance metrics (accuracy, sensitivity, precision, f_1 -score). The sample data set has imbalanced classes in the target variable as the number of instances of the negative class greatly outnumber instances of the positive class. A difficulty of using traditional classification algorithms with imbalanced data is that

they tend to classify observations as belonging to the majority class when the class of interest (positive) is represented by the minority of observations [19, 22]. This can produce a result that underestimates the occurrence of positive cases which are misclassified as false negatives. Efforts to address the issue of imbalanced data have included sampling methods such as 'boosting'. As described above, the f_1 -score is considered a better measure of performance than accuracy with data that have imbalanced classes as it takes into account both recall and precision. The study also identified features important for predicting MUPO. Tradeoffs between model complexity, performance, and interpretability are discussed.

2 METHOD

2.1 The Data

The NSDUH public data files for 2015, 2016, and 2017 were downloaded from the Substance Abuse and Mental Health Data Archive (SAMHDA) [14]. The data sets were extracted and saved as data frame objects in a python interactive notebook [23, 24]. Data from the three years ($n_1=57146$, $n_2=56897$, $n_3=56276$) were combined to create a single data set. Two outliers were identified and removed, resulting in a total sample of $N=170317$ observations (80913 male, 89404 female). According to the NSHUD codebook, the sampling design is weighted across states by population size, drawing more heavily from eight states with the largest populations (i.e., CA, FL, IL, MI, NY, OH, PA, TX), for a representative distribution that accounts for approximately 48 percent of the U.S. population. Identifying information in the NSDUH public use files is collapsed (e.g. age categories); variables related to ethnicity, immigration status, and state identifiers are removed to ensure confidentiality. The data frames were subset by column to select approximately 90 variables that included common demographic characteristics, physical health, mental health, medication usage, and illicit drug

use. Inconsistencies in the data were detected and removed with the following steps: (a) Remove missing values (i.e. NaN); (b) Recode blanks, non-responses, or legitimate skips (e.g., 99, 991, 993) to zero; (c) Recode dichotomous responses (e.g., 0=No, 1=Yes); (d) Recode ordinal variables to be consistent with amount or degree (e.g., 1=low, 2=medium, 3=high).

2.1.1 Aggregated Variables. Related features were combined to create aggregated variables. For example, a single variable indicating overall history of health problems (HEALTH) was created by combining responses for overall health (reverse scored), any previous diagnosis of STDs, hepatitis, HIV, cancer, and any hospitalization. A mental health variable (MENTHLTH) aggregated responses for adult depression, emotional distress, and suicidal thoughts or plans. Binary responses for ten of the most commonly used prescription pain medications (e.g., Hydrocodone, Oxycodone, Tramadol, Morphine, Fentanyl, Oxymorphone, Demerol, Hydro-morphone) were aggregated into a variable for any prescription opioid pain reliever use (ANYPRLUSE) in the past year. The majority of questions related to substance use had dichotomous responses that were summed to create single measures for: Tranquilizers, Sedatives, Heroin, Cocaine, and Amphetamines. Because hallucinogens varied greatly in type and potency (e.g., marijuana, psilocybin, MDMA, LSD), they were not included in the analysis. Variables for drug treatment and mental health treatment combined responses for any inpatient care, outpatient care, treatment at a clinic, emergency room visits, or hospital stays. The target variable was any prescription opioid pain reliever misuse or abuse (PRLMISEVR). The subset data frame consisted of 19 features and 170317 observations was exported to a CSV file. Four variables were highly correlated with other variables and excluded (PRLANY, PRLMISAB, HEROINUSE, HEROINFQY). Table 2 shows the list of predictor variables used for constructing the classifier models.

2.1.2 Model Construction and Evaluation. The dataset was divided using a 75 to 25 percent split to create the training set ($n_1=127738$) and testing set ($n_2=42579$). The same general procedure was used for constructing and evaluating each classification model: (i) The model was fit to the training set; (ii) New values were predicted on the holdout scores in the testing set; and (iii) Model performance on the test set was evaluated in a confusion matrix. The performance metrics of accuracy, sensitivity or recall, precision, and f_1 -score were obtained or derived from the confusion matrix. The logistic regression classifier, LDA, QDA, decision trees classifier, random forests classifier, gradient boosted trees were constructed using the caret package. The KNN classifier, support vector classifier (SVC), naive Bayes classifier, and neural network (multilayer perceptron) were built using scikit-learn.

3 RESULTS

3.1 Exploratory Data Analysis

Twenty-six percent of the total sample ($n=44596$) reported taking any pain relievers in the past year (19558 males, 25038 females). Table 4 provides the frequency and percent of respondents who reported pain reliever misuse and abuse by demographic characteristics. Approximately 11% of the sample had misused or abused prescription pain relievers at some point; this same rate of pain

Table 4: Frequency of Pain Reliever Misuse and Abuse by Demographics Features for NSDUH 2015-2017 Sample.

	PRL Misuse		No Misuse	
	N	%	N	%
Total	18237	10.7%	152080	89.3%
Male	9279	50.9%	71634	47.1%
Female	8958	49.1%	80446	52.9%
Age Group				
12-17	2262	12.4%	39315	25.9%
18-25	5577	30.6%	36476	24.0%
26-35	4370	24.0%	22251	14.6%
36-49	4173	22.0%	29571	19.4%
50+	1855	10.2%	24467	16.1%
Education Level				
School Age	2262	12.4%	39315	25.9%
Some H.S.	1849	10.1%	15334	10.1%
H.S. Grad	4134	22.7%	30223	19.9%
Some Coll.	5997	32.9%	37230	24.5%
Coll. Grad	3995	21.9%	29978	19.7%
Marital Status				
Single	6777	37.2%	64022	42.1%
Divorced	6497	35.6%	46183	30.4%
Widowed	1225	6.7%	7995	5.3%
Married	3738	20.5%	33880	22.3%

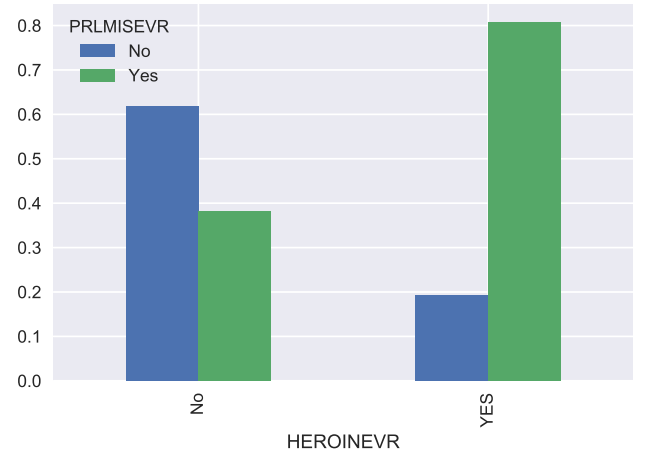


Figure 3: Proportion of Pain Reliever Misuse and Abuse as a function of Heroin Use.

reliever misuse and abuse was found in both the training set and test sets. A larger number of females reported using any prescription opioid pain relievers in the past year than males, though more males reported misusing or abusing opioid pain relievers. Pain reliever misuse was most frequent among individuals who described themselves as single, divorced, or with some college education. The proportion of pain reliever misuse and abuse was highest among individuals between 18 to 25 years of age and decreased among

Table 5: Confusion Matrices and Performance Metrics for Classification Models of Pain Reliever Misuse and Abuse

Model		Confusion Matrix		Accuracy	Sensitivity	Precision	F1-Score
K-Nearest Neighbors		No Misuse	PRL Misuse				
	No Misuse	37567	453	89.8%	0.900	0.870	0.870
	PRL Misuse	3905	654				
Logistic Regression		No Misuse	PRL Misuse				
	No Misuse	37618	3478	90.7%	0.987	0.915	0.950
	PRL Misuse	495	988				
Linear Discriminant Analysis (LDA)		No Misuse	PRL Misuse				
	No Misuse	37063	3073	90.3%	0.973	0.923	0.947
	PRL Misuse	1050	1393				
Quadratic Discriminant Analysis (QDA)		No Misuse	PRL Misuse				
	No Misuse	34786	2426	86.5%	0.913	0.935	0.924
	PRL Misuse	3327	2040				
Support Vector Machines (SVM)		No Misuse	PRL Misuse				
	No Misuse	37660	360	90.3%	0.900	0.880	0.880
	PRL Misuse	3776	783				
Naive Bayes		No Misuse	PRL Misuse				
	No Misuse	38110	4431	89.6%	0.999	0.896	0.945
	PRL Misuse	3	35				
Neural Network (MLP)		No Misuse	PRL Misuse				
	No Misuse	37488	534	90.6%	0.90	0.89	0.880
	PRL Misuse	3542	1016				
Decision Trees		No Misuse	PRL Misuse				
	No Misuse	37655	3597	90.5%	0.988	0.913	0.949
	PRL Misuse	458	869				
Random Forests		No Misuse	PRL Misuse				
	No Misuse	37556	3397	90.1%	0.985	0.917	0.950
	PRL Misuse	557	1069				
Gradient Boosted Trees		No Misuse	PRL Misuse				
	No Misuse	37955	65	89.9%	0.90	0.89	0.895
	PRL Misuse	4261	298				

Table 6: Main Parameters for Non-linear Classification Models of Pain Reliever Misuse and Abuse.

Model	Main Parameters
K-Nearest Neighbors	Number of Neighbors = 4
Support Vector Machines (RBF)	C = 10, gamma = 0.1
Naive Bayes	Cost = 0.01
Neural Network	Hidden Layers = 1
Decision Trees	Maximum Depth = 4
Random Forests	Number of Trees = 1000
Boosted Trees	Learning Rate = 0.01

older age groups. Two percent of respondents (n=3433) disclosed ever using heroin (2019 males; 1414 females). Among respondents who reported using pain relievers in the past year, the rate of misuse and abuse of pain relievers was twice as large for individuals who

reported using heroin than those who had not used heroin (shown in Figure 3), which is consistent with past findings that indicate a connection between MUPO and heroin use [7, 8].

3.2 Classifier Model Performance

Performance of the classification models was evaluated in the confusion matrices reported in Table 5, which includes the metrics of accuracy, sensitivity (recall), precision, and f_1 -score. The model accuracy scores ranged from 86.5% to 90.7%. Because of the unbalanced classes in the sample data—the proportion of respondents who had not misused or abused pain relievers (89%) was much greater than the proportion who had—the f_1 -score was used as the preferred performance metric rather than accuracy. The f_1 -scores ranged from 0.87 to 0.95. Logistic regression and the random forest model were tied for the highest f_1 -score (0.95) followed by the decision tree model (0.949). The random forests model identified more true positives (1069) and fewer false negatives (3397) than

the single decision tree or logistic regression. QDA identified the highest number of true positives, but detected far fewer true negatives than logistic regression or LDA. In general, default parameter setting were used for most models; the main parameter settings for the classification models are presented in Table 6. The results for the three top performing models are described below.

3.2.1 Logistic Regression. The parameter estimates and odds ratios for the logistic regression model are presented in Table 7. The coefficient estimates were all statistically significant with the exception of mental health treatment. The model was rerun leaving out the non-significant variable, but the performance metrics did not change and the full model with all of the predictors was retained. The coefficient estimates ($\beta_1 \dots \beta_j$) are interpreted as the change in the log of the odds of the dependent variable (PRLMISAB) occurring given a one unit change in each independent variable ($X_1 \dots X_j$), holding constant the effects of other independent variables. The parameters relate to the log of the odds ratio rather than to the dependent variable directly. The odds are calculated as the probability of the event occurring divided by the probability of the event not occurring ($\frac{P}{P-1}$). The odds ratio is obtained by taking the antilog of the estimated coefficient, which is the exponentiated parameter estimate (e^x). For an interval scaled independent variable, the odds ratio is interpreted as the multiplicative increase in the odds of the dependent variable event happening given a one unit change in the value of the independent variable, assuming the effects of all other independent variables are held constant. For a dichotomous variable (i.e., sex), it represents the multiplicative increase in the odds of the dependent variable event happening if the event represented by the independent variable occurs. In general, taking the antilog of the coefficient estimate, subtracting 1 from it, and multiplying the result by 100, provides the percent change in the odds for a unit increase in the independent variable [16].

For a one unit increase in cocaine use, the odds of misusing or abusing pain relievers increased by 1.96 times or 96%, holding the effects of all other variables constant. For a one unit increase in amphetamine use, the odds of pain reliever misuse and abuse (PRLMISAB) increased by 1.81 times or 81%. A one unit increase in tranquilizer use was associated with a 1.48 or 48% increase in the odds of PRLMISAB. For a unit increase in mental health issues, the odds for PRLMISAB increased by 1.14 or 14%. For a one unit increase in heroin use, the odds of PRLMISAB increased by 2.12 times or 112%. For a one unit increase in age category, the odds of PRLMISAB decreased by 1.20 times or 20%. A one unit increase in employment is associated with a 1.21 or 21% increase in the odds of PRLMISAB. For a unit increase in substance treatment, the odds for PRLMISAB increased by 1.23 or 23%. For a one unit increase in sedative use, the odds of PRLMISAB increased by 1.30 times or 30%. In terms of sex, the odds of misusing or abusing pain relievers for females decreased by 1.17 times compared to males; in other words, the odds of PRLMISAB for females was 17% lower than for males. The odds ratio indicates the ratio of the positive event to the negative event, but does not reveal the probability of the positive outcome. The terms in the logistic regression equation can be rearranged to obtain the function of the event probability, as presented in equation 3:

Table 7: Parameter Estimates for Logistic Regression Model and Log-odds Ratios for Pain Reliever Misuse and Abuse (Training Set).

Predictor	Estimate	Std. Err.	Z-Value	Odds Ratio
(Intercept)	-1.901	0.307	-6.19 ***	
Cocaine	0.672	0.015	44.43 ***	1.96
Amphetamines	0.592	0.017	35.55 ***	1.81
Tranquilizers	0.393	0.011	34.57 ***	1.48
Mental Health	0.133	0.005	26.96 ***	1.14
Heroin use	0.750	0.041	18.25 ***	2.12
Age Category	-0.185	0.011	-17.28 ***	1.20
Employment	0.191	0.013	14.60 ***	1.21
Treatment	0.203	0.015	13.57 ***	1.23
Sedatives	0.264	0.024	10.98 ***	1.30
Education	0.105	0.010	10.12 ***	1.11
Health	0.093	0.010	9.15 ***	1.10
Sex	-0.153	0.021	-7.28 ***	1.17
Married	0.052	0.010	5.42 ***	1.05
Year	-0.071	0.020	-3.58 ***	1.07
City/Metro	-0.040	0.013	-3.08 **	1.04
MH Treatment	-0.001	0.014	-0.07	1.00
Note: p-value	< 0.001***		< 0.01**	

$$P = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_p * X_p)]} \quad (3)$$

3.2.2 Decision Trees. The decision tree model was pruned to a maximum depth of 4, which means the algorithm split on four consecutive features (see Figure 4). The algorithm selected cocaine as the root node which includes the n=127738 observations in the training set. With a classification tree, we predict that each observation will belong to the class of training observations in the region to which it belongs, and want to determine not only the class prediction belonging to a terminal node region, but also the class proportions among the training observations in that region [9]. One way to interpret a decision tree is by following the number of samples represented at the split for each node. Another way to interpret the decision tree in Figure 4 is by the examining the proportion of observations of class A captured by that leaf over the entire number of observations captured by the leaf during model training. Starting from the root node, the branch to the left represents n=112730 observations with no or low cocaine use; of those, 0.07 or 7% reported misusing or abusing opioid pain relievers. Following the right branch are n=15008 observations positive for cocaine use, of which 0.39 or 39% reported pain reliever misuse or abuse. This means that the proportion of pain reliever misuse and abuse was more than five times as large for individuals who reported using cocaine than those who had not.

The second split was based on heroin use; the left branch included n=12876 respondents who had not used heroin, of which 0.34 or 34% reported pain reliever misuse or abuse. Following the right branch to the terminal node (i.e., 'leaf'), n=2132 individuals reported heroin use, of which 0.70 or 70% had misused or abused pain relievers. Thus, the proportion of PRLMISAB was twice as large for respondents

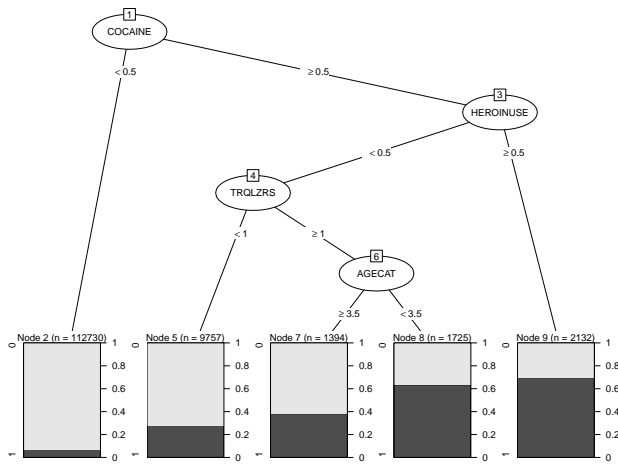


Figure 4: Decision Tree Model of Pain Reliever Misuse and Abuse fit to the Training Set.

who had used heroin than those who had not used heroin (as seen in Figure 3). The third split in the decision tree was based on tranquilizers. The left branch represented $n=9757$ individuals who reported no or low tranquilizer use; of these, the proportion of PRLMISAB was 0.28 or 28%. The branch to the right represented $n=3119$ observations with moderate to high tranquilizer use, of which 0.52 or 52% reported PRLMISAB. The rate of pain reliever misuse and abuse for individuals with moderate to high tranquilizer use was almost twice as large as for those reporting no to low tranquilizer use. The fourth split was based on age category; the branch to the left represented $n=1394$ individuals age 36 or older, of which 0.38 or 38% reported PRLMISAB. The branch to the right represented $n=1725$ individuals age 35 and younger, of whom 0.63 or 63% reported PRLMISAB. This findings shows that pain reliever misuse and abuse was much more likely among respondents age 35 or younger than among individuals older than 35 years.

3.2.3 Feature Importance for the Random Forests Model. The random forest model was fit using 1000 trees, with all of the features considered at each node to determine the randomness of each tree. After a large number of trees is generated, each tree represents a vote for the most popular class. Although random forests perform well with data that has imbalanced classes, it can be difficult to interpret the result of the averaged trees. Feature importance is a model summary for random forests that rates how important each feature is for classification decisions made in the algorithm. The Gini index provides a measure of node purity which is used to evaluate the quality of a particular split; a small value indicates that a node predominantly contains observations from a single class. Feature importance was measured by the mean decrease in Gini coefficient (Table 8), which indicates how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. For classification trees, the splits are chosen so as to minimize entropy or Gini impurity in the resulting subsets. For random forests, variables that result in nodes with higher purity have a higher decrease in the Gini coefficient. Feature importance

Table 8: Feature Importance for the Random Forest Model.

Predictor	Mean Decrease in Gini Score
Cocaine	1990.45
Amphetamines	1340.20
Mental Health	1314.92
Health	1015.29
Tranquilizers	978.70
Age Group	899.45
Education	893.39
Heroin Use	780.27
City/Metro	664.99
MH Treatment	615.17
Married	565.02
Employment	564.11
Year	472.63
Sex	392.41
Treatment	388.30
Sedatives	223.47

is computed by aggregating the feature importance over trees in the random forest, and gives non-zero importance to more features than a single tree. A feature may have a low importance value because another feature encodes the same information. Table 8 provides the feature importance for the random forests model sorted by the mean decrease in the Gini coefficient. The algorithm selected cocaine as the most informative feature for predicting pain reliever misuse and abuse. In contrast to the single decision tree, amphetamines, mental health, and health were selected among the top four most important features in the random forests model, followed by tranquilizers, age category, and heroin use, which were ranked as more influential in the single tree (Figure 4).

4 DISCUSSION

The results show that there were imbalanced classes in the sample data given that the proportion of respondents who reported no MUPO (i.e., negative class) was much larger than the 11 percent of respondents who reported pain reliever misuse and abuse. Logistic regression, random forests, and decision trees performed better than other classifier models using the f_1 -score as the appropriate performance metric. Traditional algorithms tend to classify new observations based on the majority class, as seen with the naive Bayes model, which had a relatively high f_1 -score despite identifying only 35 true positive instances and having a high number of false negatives. The random forests model performed slightly better than the other models as it correctly predicted more true positive instances than the logistic regression and had fewer false negative errors than the single decision tree. The QDA and LDA models both identified the largest number of true positives, but each correctly labeled fewer true negatives compared to the logistic regression model, which had higher f_1 -score and accuracy overall. This finding indicates that the decision boundary for classifying pain reliever misuse and abuse can be modeled as a linear function of the predictor variables [9, 12]. The decision tree and random

forests models would be preferred if pain reliever misuse and abuse were modeled as a non-linear function of the predictors,

Cocaine use was selected by all three models as the most informative predictor of pain reliever misuse and abuse (PRLMISAB). The odds of PRLMISAB increased 99 percent given a one unit increase in cocaine use, holding other independent variables constant. In terms of frequency, the proportion of PRLMISAB among individuals who reported moderate to high cocaine use was more than five times as large as for individuals who reported no or low cocaine use. Logistic regression and random forests both identified amphetamine use as the second most important feature for predicting PRLMISAB. The odds of PRLMISAB increased by 84 percent for a unit increase in amphetamine use. In contrast, the single decision tree identified heroin use as the second most important predictor. Of those respondents who disclosed heroin use, more than two-thirds reported misusing or abusing pain relievers. In other words, the proportion of PRLMISAB was twice as large for respondents who had used heroin as for those who had not. Similarly, logistic regression model revealed that the odds of PRLMISAB increased by 112 percent given previous heroin use, holding other variables constant. These findings are consistent with past studies that indicate a connection between pain reliever misuse and abuse and heroin use [5–8]. Tranquilizers were identified as the third most important variable by both the decision tree and logistic regression models. The odds of PRLMISAB increased 1.46 times for a one unit increase in tranquilizer use. Similar to the pattern observed for heroin, the proportion of PRLMISAB was nearly twice as large for respondents with moderate to high tranquilizer use as for individuals with no to low tranquilizer use.

Age category, mental health, health issues, education level, and previous drug or alcohol treatment were also identified as important features for predicting PRLMISAB; however, the models differed in the importance they assigned to these variables. A general finding related to age is that pain reliever misuse and abuse was more prevalent among respondents between 18 to 25 years of age and decreased among older individuals. Although more females reported using pain relievers than males, more males reported misusing or abusing pain relievers than females, and the odds of pain reliever misuse or abuse was 17 percent lower for females as for males.

4.1 Model Complexity and Interpretability

Every classification model has advantages and limitations. There is a tradeoff between model complexity, performance, and interpretability. Simple models provide interpretable solutions but have lower accuracy, whereas complex models yield improved performance but are more difficult to interpret. The advantage of logistic regression is that it provides coefficient estimates for individual features that can be interpreted in terms of odds ratios. The predicted outcome (\hat{Y}) is represented by the weighted combination of all the independent variables. The logit function is linear in the parameters, but the probabilities of the expected outcome are non-linear. The drawbacks of logistic regression are that it does not perform well for modeling non-linear relationships between the target outcome and predictor variables, or with high dimensional data (e.g., $p > n$).

Decision trees are useful for modeling nonlinear data and can be interpreted in terms of the frequency or proportion of observations selected at each branch. A single tree visually represents the decision process in a manner that is effective for communicating results; however, a disadvantage is that decision can become very complex without pruning, especially as the number of predictors increases. Limiting the decision tree to a depth of four nodes improves the interpretability, but reduces the generalizability of the model to new observations. Random forests reduce overfitting by averaging multiple different trees to identify class membership, which improves generalizability to new data. A limitation of random forests is that feature importance is determined by the mean decrease in Gini score, a measure of node purity that is difficult to interpret in terms of the outcome. Gradient boosting typically improves accuracy using many simple models iteratively and correcting for individual trees; however, in the present study, the boosting model did not perform better than random forests.

As the complexity of a model increases, interpretation becomes more difficult, as seen with neural networks. The multilayer perceptron (MLP) is one of the most widely used neural network models for classification. With a limited number of predictor variables and single hidden layer, it is possible to interpret the relations among weights and nodes in the hidden layer (see Appendix). As the number of predictors and hidden layers increases, complex neural network become opaque to interpretation and represent a “black box” model that is not comprehensible at a human level. In applying sophisticated algorithms in predictive modeling, interpretability is often sacrificed for greater model accuracy [25].

4.2 Limitations

A surprising finding is that several models which typically perform well for binary classification tasks (gradient boosting, SVM, neural networks) did not perform better than was observed in this study. This may be due to imbalanced classes in the sample data. A limitation of the study is that the default settings were used in constructing the classifier models. Many complex models are sensitive to parameter settings and scaling of the data. Additional parameter tuning could have improved performance of complex models. Cross-validation can be used to select hyperparameters that optimize model performance. Another limitation is that the aggregated features represent a subset of the entire range of features in the NSDUH datasets. In a future study, it may be useful to include a more comprehensive set of features to identify additional variables for predicting opioid dependence and addiction. Given the large proportion of the sample that had not previously used any prescription opioid pain relievers, it may be useful to classify pain reliever misuse and abuse using a subset of individuals who have previously used pain medications, although this would result in a reduced sample. It is a widely accepted truism in predictive analytics that, “more data is always better”.

4.3 Opioid Addiction

Although drug addiction has many similar characteristics to other chronic medical illnesses, there are unique challenges to the treatment of addiction. According to a classical conditioning theory of addiction, situational cues or events can elicit a motivational state

underlying the relapse to drug use [26]. Following treatment, many addicted individuals return to the same environments associated with their drug use. Addictive behavior can be reinstated by exposure to drug-related cues or stressors in the environment, putting individuals in recovery at risk for relapse and possible overdose. In the case of prescription opioids, anyone is potentially at risk for misusing opioids and becoming addicted to pain relievers. Rather than labeling people as ‘addicted’ or ‘not addicted’, it may be useful to consider people as more or less susceptible to misusing or abusing opioid pain medication. This study also provides further support for the connection between pain reliever misuse and heroin use. The rate of pain reliever misuse and abuse was twice as great for respondents who reported using heroin as for those who had not. Recent statistics from the CDC show that the use of illicit and synthetic opioids is a major contributing factor in the increase of opioid overdose deaths [3]. If the crisis in opioid addiction is a true epidemic, network analysis may be useful for describing the spread or diffusion of drug use and addictive behavior within social networks.

5 CONCLUSION

Predictive modeling offers several useful approaches for analyzing the misuse and abuse of prescription opioids and identifying factors that contribute to MUPO. This study compared ten classification models of pain reliever misuse and abuse using the f_1 -score to evaluate model performance with unbalanced classes in the data. Logistic regression, random forests, and decision trees had the best performance compared to other models. Cocaine use was selected as the most informative variable by all three models, followed by amphetamine use which was selected as the second most important feature by the logistic regression and random forests models. The models differed in the importance they assigned to heroin use, tranquilizers, and demographic features such as age group, mental health, and health problems. A general conclusion is that there are tradeoffs between model complexity, performance, and interpretability. A simple decision tree provides an interpretable model that is prone to overfitting. Random forests reduce overfitting and improve generalizability, but are difficult to interpret. Logistic regression strikes a balance between complexity and interpretability, but does not perform well with non-linear relationships or high-dimensional data. Additional research is needed to understand the relationships among the variables identified by predictive models. The findings may inform decision making and policy efforts to address the opioid crisis and reduce the risk of overdose death.

ACKNOWLEDGMENTS

Portions of this paper were completed as part of a course project for ‘Big Data Applications and Analytics’ taught by Professor Gregor von Laszewski at Indiana University in Fall 2017. Thanks to the teaching assistants Juliette Zurick and Miao Jiang. Thanks to Dallas J. Elgin for encouragement and Ed Miles for helpful comments.

REFERENCES

- [1] Nora D. Volkow, Thomas R. Frieden, Pamela S. Hyde, and Stephen S. Cha. Medication-assisted therapies: Tackling the opioid-overdose epidemic. *New England Journal of Medicine*, 370(22):2063–2066, 2014. PMID: 24758595.
- [2] National Institute on Drug Abuse. Opioid overdose crisis. online, Jan 2018.
- [3] Centers for Disease Control and Prevention. Understanding the epidemic: Opioid overdose deaths. online, Oct 2018.
- [4] Rose A. Rudd, Noah Aleshire, Jon E. Zibbell, and R. Matthew Gladden. Increases in drug and opioid-involved overdose deaths in the United States, 2010–2015. *MMWR Morbidity and Mortality Weekly Report*, 65(16):420–423, April 2016.
- [5] C. M. Jones, J. Logan, M. Gladden, and M.K. Bohm. Vital signs: Demographic and substance use trends among heroin users. *MMWR Morbidity and Mortality Weekly Report*, 64(26):719f–725, July 2015. Published online.
- [6] C. M. Jones. Heroin use and heroin use risk behaviors among nonmedical users of prescription opioid pain relievers - United States, 2002–2004 and 2008–2010. *Drug and Alcohol Dependence*, 132(1–2):95–100, September 2013.
- [7] P K Muhuri, J C Gfroerer, and M C Davies. Associations of nonmedical pain reliever use and initiation of heroin use in the United States. *Center for Behavioral Health and Statistics Quality (CBHSQ) Data Review*, Aug 2013.
- [8] G.J. Unick, D. Rosenblum, S. Mars, and D. Ciccarone. Intertwined epidemics: National demographic trends in hospitalizations for heroin- and opioid-related overdoses, 1993–2009. *PLoS ONE*, 8(2):e54496, 2013.
- [9] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, New York, NY, 2013.
- [10] Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*. Springer, New York, NY, 2013.
- [11] Andreas C. Muller and Sarah Guido. *Introduction to Machine Learning*. O’Reilly, Sebastopol, CA, 2017.
- [12] Sebastian Raschka and Vahid Mirjalili. *Python Machine Learning*. Packt, Birmingham, UK, 2017.
- [13] M. Herland, T. M. Khoshgoftar, and R. Wald. A review of data mining using big data in health informatics. *Journal Of Big Data*, 1(2), 2014.
- [14] Substance Abuse, Center for Behavioral Health Statistics Mental Health Services Administration, and Quality. National survey on drug use and health (nsduh) 2015. Online data archive, United States Department of Health and Human Services., Ann Arbor, MI, 2016.
- [15] George E. P. Box, J.S. Hunter, and W.G. Hunter. *Statistics for Experimenters (2 edition)*. John Wiley and Sons, John Wiley Sons, 2005.
- [16] Damodar N. Gujarati and Dawn C. Porter. *Basic Econometrics, 5th edition*. McGraw-Hill Irwin, 2009.
- [17] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
- [18] Wikipedia. Evaluation of binary classifiers. online, Oct 2018.
- [19] S. Yun, A.K.C. Wong, and M.S. Kamel. Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719, 2009.
- [20] Rice J.B., White A.G., Birnbaum H.G., Schiller M., Brown D.A., and Roland C.L. A model to identify patients at risk for prescription opioid abuse, dependence, and misuse. *Pain Medicine*, 13(9):1162f–1173, September 2012.
- [21] S.E. McCabe, B.T. West, C.J. Teter, and C.J. Boyd. Medical and nonmedical use of prescription opioids among high school seniors in the United States. *Archives of Pediatric Adolescent Medicine*, 166(9):797f–802, 2012.
- [22] I. Brown and C. Mues. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 39(3):3446–3453, 2012.
- [23] Wes McKinney. *Python for Data Analysis*. O’Reilly Media Inc., Sebastopol, CA, 2017.
- [24] Jake VanderPlas. *Python Data Science Handbook*. O’Reilly Media Inc., Sebastopol, CA, 2017.
- [25] D.J. Elgin. Utilizing predictive modeling to enhance policy and practice through improved identification of at-risk clients: Predicting permanency for foster children. *Children and Youth Services Review*, 91:156–167, Aug 2018.
- [26] Yavin Shaham, Uri Shalev, Lin Lu, Harriet de Wit, and Jane Stewart. The reinstatement model of drug relapse: history, methodology and major findings. *Psychopharmacology*, 168(1):3–20, Jul 2003.

A APPENDIX

A.1 Supplemental Figure

A.1.1 Neural Network. The multilayer perceptron (MLP) is a simple back-propagation neural network that takes the features as input; a single hidden layer comprises the weighted combination of the input variables, and the output layer represents a response probability. During model training, the weights of the predictor variables are first randomly initialized and then iteratively adjusted to minimize an error function (e.g. gradient descent) [22].

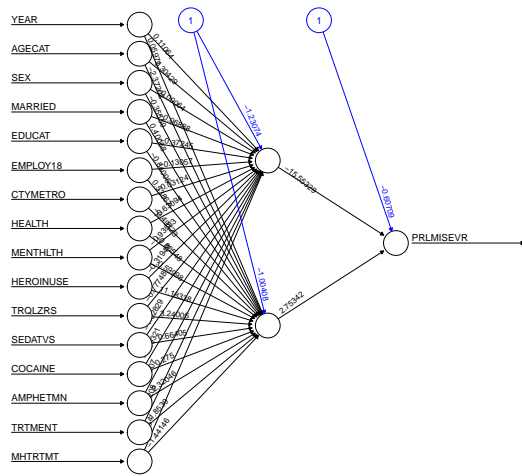


Figure 5: Neural Net Classifier: Multilayer Perceptron with a Single Hidden Layer.