# Machine learning to predict factors contributing to treatment success for drugs / alcohol substance use disorder

by

**Muriel Banze**

A Project Report Submitted
in
Partial Fulfillment of the
Requirements for the Degree of
Master of Science
Supervised by

Dr. Michael McQuaid

School of Information

B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York

August 2021

The project report “Machine learning to predict factors contributing to treatment success for drugs / alcohol substance use disorder” by Muriel Banze has been examined and approved by the following Examination Committee:

Dr. Michael McQuaid
Senior Lecturer
Project Committee Chair

Prof. David Patric
Lecturer
Project Committee

# Abstract

## Machine learning to predict factors contributing to treatment success for drugs / alcohol substance use disorder

**Muriel Banze**

**Supervising Professor: Dr. Michael McQuaid**

Substance abuse such as drugs or alcohol has always been an issue that has had a significant impact on an individual's health and maintaining sobriety following treatment has always been challenging. Moderate drug use can lead to addiction, which can be hard to quit. Symptoms could vary from mild to extremely addictive, and drug overdosing can lead to a variety of long-term chronic illnesses. A major problem occurs when an individual has been abstinent for a long period and seeks treatment: they are more than likely to relapse at least once. This is when they lose self-control and may end up overdosing, which is extremely dangerous and can lead to death. Machine learning was used to identify the factor responsible for drug addiction and to determine the number of days rehab consultation treatment required for patient to fully recover. To identify potential factors that contribute to effective treatment, this project developed supervised machine learning models that were used to implement classification algorithms such as Random Forest. When compared to other models such as Logistic Regression, Naive Bayes, K-Nearest Neighbors, and Decision Trees, the Random Forest model performed the best in terms of accuracy, sensitivity, specificity, ROC, and AUC scores. This algorithm was fine-tuned further to improve the model's accuracy. This model is then used to identify factors such as duration of stay, drug type, frequency of use, and type of services required to generate predictions for a patient to successfully finish their treatment.

Keywords: Substance abuse, machine learning, treatment drug addiction, treatment success

# Contents

# 1 Introduction

Substance abuse is a mental illness characterized by compulsive drug seeking that has serious health consequences. Repeated use of illegal substances can lead to cognitive impairment and physical deterioration, both of which could adversely impact daily activities. According to the Centers for Disease Control and Prevention (CDC), about 564,000 individuals have died of opioid overdose since 1999 to 2020. In the USA (for disease control & prevention 2022, 2022), over 70 percent of all overdose deaths were due to opioid overdose were reported in 2019. As per recent reports during the Covid-19 pandemic, patients with substance use disorder (SUD) were more likely to suffer from kidney failure, type 2 diabetes, lung disease and cancer(Wang et al., 2021). These patients had worse outcomes mostly resulting in death when compared to other patients without SUD.

The most used substances are nicotine, tobacco, and alcohol. Marijuana and heroin are drugs that mimic the neurotransmitter in the brain, which is also known as the chemical messenger that controls muscular action. These drugs can produce abnormally large amounts of signals, resulting in the 'high' effect. Most of these substances influence the brain's reward system causing euphoria. People usually begin to consume drugs on a voluntary basis and addiction can develop if not managed. These individuals may be affected by outside factors such as environmental circumstances, life, and genetics. As a person continues to use these substances, the brain adjusts to the overwhelming surges of dopamine, making it difficult to control since the person may no longer find pleasure in the things that brought them joy previously (Pitchers et al., 2018). As a result, many revert to drug abuse to restore dopamine levels to normal. Because of this, people continue to take higher doses as their tolerance to these drugs develops. Effects of these abuse can lead to depression, muscle breakdown, gum disease, cardiac arrest, kidney damage and failure. Several treatments exist that cater to the need of the patients (Juergens & Hampton, 2021). Patients in the inpatient program are advised to stay in a controlled environment. This service is also known as residential treatment. Patients stay in these facilities to recover from addiction. These patients stay in the clinic for about 24 hours and receive clinical and emotional support. Outpatient treatment entails devoting more than 10-12 hours per week for a set number of weeks. This treatment can last anywhere from three months to more than a year.

Discontinuing of drugs is only one part of the problem; the real issue arises during the recovery process (on Drug Abuse, 2020). The side effects of drugs have had a significant impact on people's health, as well as their work and family lives. Addiction does not have a cure but it can be successfully controlled. Addiction can have a significant impact on an individual's life; therefore, the treatments provided to patients should address the patient's needs. The patient can recover with the help of family, counselors, medicines, and positive social influence. Relapse after treatment is quite common amongst individuals with SUD and there is a need for a treatment that is long

term effective. Repeated admissions to treatment facilities due to relapse can occur in patients that have just completed medical detoxification. Accounting to each individual patient characteristic, it is necessary to provide them with the right kind of treatment. By understanding the basic needs of individuals based on the severity of their addiction, we can administer the right kind of treatment to avoid relapse.

## 2 Related Work

Machine learning has been used to detect substance abuse, assess hazards, and predict treatment effectiveness. Nath et al., 2017 presented a methodology for determining whether or not a person has consumed volatile substances. Artificial neural networks were employed and type is the Feed forward back propagation neural network, done by creating two artificial neural network modules namely ANN-C and ANN-D, with ANN-D predicting volatile drug consumption and ANN-C determining time of usage for day, week, month, and year. According to the findings of this study, drug user accuracy was 81.1 percent (ANN-D) and time of usage was 71.1 percent (ANN-C). Personality characteristics, impulsivity, demography, and sensation were all included in this study. The drawback to this study is that just five factors are not enough attributes to provide a high accuracy in the ANN-D and ANN-C model.

Afzali et al. (2019) describes adolescent usage of alcohol by making a cross cultural evaluation between Canada and Australia. Using machine learning it compared using binary logistic classification along with RF, SVM, lasso. The drawback of this study was that personality traits were not considered since it can be a factor contributing to consuming alcohol. Wetherill et al. (2019) used SVM classification models of rsFC data to identify smoking addiction in people who matched non-smokers controls. These people were discovered to have lower functional connectivity measurements. The goal of this study is to identify biological flaws in smokers. Boslett et al. (2020) aimed to provide opioid overdose using statistical analysis and concluded that Logistic regression outperformed RF. Accuracy of the models increases by including contributing death factors due to opioid usage.

Chawla et al. (2002), defines various approaches to dealing with imbalanced classes. The SMOTE technique for synthetic data generation is examined to determine the difference in model performance. This study also suggests that using both SMOTE and under sampling approaches resulted in higher prediction accuracies. Feature importance as mentioned by Doyen et al. (2021), offers metrics to investigate the importance of features. This study implemented the hollow tree super and suggests that it can be incorporated in the machine learning models. This feature provides easy interpretation of data and metric. According to Kamp et al., 2019, upon treatment completion, studies show that there was a decrease in craving and psychiatric symptoms. This study also showed that cognitive flexibility and processing speed amongst patients with methamphetamine abuse improved significantly. Length of stay and treatment completion has

also proven to be an essential factor as per Zarkin et al. (2002) since it shows decrease in signs of criminal activity as well as unemployment for a person under SUD.

Based on the study by Tapia-Galisteo et al. (2020), talks about implement machine learning algorithm with large high dimensional heterogeneous data. This study focused its results on predicting cocaine disorder for patients seeking inpatient type of treatment. The SVM machine learning was used in this study, and the results indicated that lowering the drop out rate of patients would lead to better resource management. According to Thompson et al. (2020) research, suit consultations involve integrating peer-based recovery strategies to reduce readmission rates. Model tuning and overfitting as per Kuhn and Johnson (2013) suggests techniques such as resampling and data splitting. Final tuning parameters are used to obtain the optimal result. As per James et al. (2013), decision trees are pruned using bagging and bootstrapping. Based on the findings of this study, we can conclude that as out of bag error is reduced, overall accuracy improves. According to Kelkar and Bakal (2020), parameter tuning for the random forest algorithm improved accuracy and the Cohen's Kappa value. Because Cohen's value is affected, the maximum depth of the tree should be high. The tuneRanger package for random forest was another model suggested by Probst et al. (2019) this study that provided a framework for model based optimization. Another study Nicodemus et al. (2010) based on permutation based important features for random forest have been made and how they can be synthesized to get optimal results.

Based on these findings, we may conclude that machine learning algorithms can assist in predicting the success factors for a treatment. Patient factors such as service type and drug usage help in understanding the right type of treatment to be administered. These studies also suggest the need for parameter tuning to improve the model accuracy and prediction performance. Studies have also suggested the usage of synthetic data to manage unbalanced classes. As per the studies mentioned above Logistic Regression offered the most accuracy of the models developed in the majority of cases. These studies also show that staying at the institution for a longer amount of time and completing treatment all contribute to a better recovery and lessen the need to relapse.

# 3 Goals and Objectives

## 3.1 Problem / Purpose:

In an ideal scenario, a person with a substance abuse disorder would be cared for properly and would not relapse. In reality, most patients do not receive the right services and are released early due to poor diagnosis. As a result, there is a need to discover factors that can help to a healthy recovery once the patient's treatment is completed. Machine learning can be used to estimate the chances of a patient recovering from substance abuse depending on the various services provided at the facility.

## 3.2 Goals and Objectives:

The overall goal of this project is to suggest factors that can aid in successful treatment. This project will indicate the number of days for a patient to fully recover.

### 3.2.1 Goal 1: Data pre-processing:

Objective 1: Analyze the dataset from the US Department of Health and Human Services. Objective 2: Carry out data pre-processing tasks such as handling null values and feature selection. Objective 3: Split the dataset into train and test sets and check for imbalanced data.

### 3.2.2 Goal 2: Perform Exploratory Data analysis:

Objective 1: Build machine learning algorithms such as Logistic regression, K-Nearest Neighbor, Decision trees, Naive Bayes, and Random Forest. Objective 2: Select the best performing model based on accuracy, precision, recall, specificity, ROC (Receiver Operating Characteristics) curve and AUC values. Objective 3: Based on various input data, make predictions to identify essential factors for successful treatment completion.

### 3.2.3 Goal 3: Data visualization:

Objective 1: Create visualizations based upon the frequency distributions for the features using bar charts and show the best factors that are crucial for treatment completion.

# 4 Methodology

## 4.1 Data source and description

The dataset was obtained from the website for Substance Abuse and Mental Health Services Administration SAMHDA of the United States Department of Health and Human Services SAMHDA, 2021. The Treatment Episode Data Set (TEDS - D) contains national level data which is collected from the treatment facilities for substance use. It contains detailed level patient data who were discharged from the facility in the year 2019. This dataset comprises of about 1,722,503 rows along with 76 columns each describing patient data with regards to patient characteristics such as age, race, gender as well as type of services given at the rehabilitation facility. To implement our model, only a subset of data is used. This subset consists of about 193,267 rows of data with 30 columns. All field records are encoded into numeric values ranging from 1 to 37, with missing values marked as $-9$. Tables 1-4 show the variable descriptions utilized in the analysis for each variable.

**Independent variables:** A total of 29 variables were chosen as the independent variables that were to be included in the analysis. These predictors included patient details such as age, education, gender, race, ethnicity, marital status, employment status, living arrangements, and veteran status, treatment characteristics such as length of stay, medicated assisted opioid therapy and services. Other patient characteristics such as substance use type, primary source of referral, and mental health disorder. The case-id has also been included since it can provide details for an individual. Substance use details such as the type of substance (alcohol, marijuana etc.), route of administration, frequency with regards to age of first use.

**Dependent Variable (Target):** The variable labeled 'Reason' is the response variable that will aid in determining whether or not the given patient has completed treatment. This variable takes on seven values, each of which indicates the reason for discharge. However, all fields except "Treatment Completed" will be recorded as 0.

Exclusion criteria: All fields that contained missing data were not included in the analysis. Fields containing data which indicated data at the time of discharge were also excluded from the analysis since it did not contain much information. Other information relating to state codes and mode of payment were not included as well.

**Data visualizations** The dataset consists of categorical variables and with the use of plots we can clearly see the frequency distribution for each feature. The library such as ggplot2 was loaded which can be used to create visually pleasing charts. 0 – Treatment Incomplete, 1 – Treatment Complete.

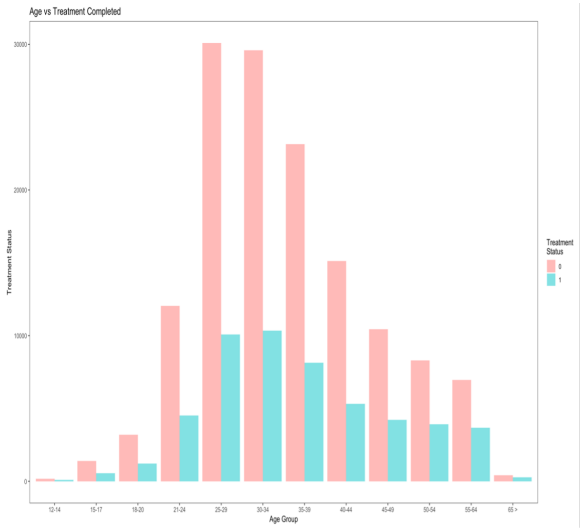


Figure 1. Bar chart comparing Age Group vs Treatment Status, (0:Treatment Incomplete, 1:Treatment Complete)

Figure 1 and Figure 3 depict characteristics such as age, and the type of substance ingested. Figure 1 depicts the age range of 12 to 65 years. We can see from this graph that those between the ages of 25 and 39 are more likely to complete their therapy.

Table 1. Description of dataset variables (Part a)

| Variable | Description | Missing Values | Encoded Value |
|---|---|---|---|
| AGE | Age recorded at the time of admission | None | 1 (12–14 years)<br>2 (15–17 years )<br>3 (18–20 years)<br>4 (21–24 years)<br>5 (25–29 years)<br>6 (30–34 years)<br>7 (35–39 years)<br>8 (40–44 years)<br>9 (45–49 years)<br>10 (50–54 years)<br>11 (55–64 years)<br>12 (65–95 years) |
| ALCDRUG | Type of substance - alcohol / other drugs | None | 0 - None<br>1 - Alcohol only<br>2 - Other drugs only<br>3 - Alcohol and other drugs |
| CASEID | Case id number | None | N/A |
| DSMCRIT | DSM diagnosis | 403,507 | 1 - Alcohol-induced disorder<br>2 - Substance-induced disorder<br>3 - Alcohol intoxication<br>4 - Alcohol dependence<br>5 - Opioid dependence<br>6 - Cocaine dependence<br>7 - Cannabis dependence<br>8 - Other substance dependence<br>9 - Alcohol abuse<br>10 - Cannabis abuse |
| EDUC | Education | 145,388 | 1 - No school/ kg – grade 8<br>2 - Grades 9 – 11<br>3 - Grade 12/(GED)<br>4 - 1-3 years of university/college<br>5 - 4 years of college/BA/BS/university |
| EMPLOY | Employment status | 130,864 | 1 – Full time<br>2 – Part time<br>3 – Unemployed<br>4 – Not in labor force |
| ETHNIC | Ethnicity | 58,922 | 1 - Puerto Rican<br>2 - Mexican<br>3 - Cuban<br>4 - Not of Hispanic or Latino origin<br>5 - Hispanic or Latino |

Table 2. Description of dataset variables (Part b)

| Variable | Description | Missing Values | Encoded Value |
|---|---|---|---|
| SERVICES | Treatment type | None | 1 – Detox, 24-hour hospital inpatient<br>2 - Detox, 24-hour hospital free standing residential<br>3 – Rehab/residential (non-detox)<br>4 - Rehab/residential, (<30 days)<br>5 - Rehab/residential, (>30 days)<br>6 – Ambulatory, intensive outpatient<br>7 – Ambulatory, non-intensive outpatient<br>8 – Ambulatory, detoxification |
| FREQ1 | Frequency of use at admission (primary) | 201,346 | 1 - No use<br>2 - Some use<br>3 - Daily use |
| FREQ2 | Frequency of use at admission (secondary) | 780,278 | 1 - No use<br>2 - Some use<br>3 - Daily use |
| FREQ3 | Frequency of use at admission (tertiary) | 1,325,778 | 1 - No use<br>2 - Some use<br>3 - Daily use |
| FRSTUSE1 | Age at first use (primary) | 120,542 | 1 - 11 years and under<br>2 - 12–14 years<br>3 - 15–17 years<br>4 - 18–20 years<br>5 - 21–24 years<br>6 - 25–29 years<br>7 - 30–95 years |
| FRSTUSE2 | Age at first use (secondary) | 782,956 | 1 - 11 years and under<br>2 - 12–14 years<br>3 - 15–17 years<br>4 - 18–20 years<br>5 - 21–24 years<br>6 - 25–29 years<br>7 - 30–95 years |
| FRSTUSE3 | Age at first use (tertiary) | 1,308,386 | 1 - 11 years and under<br>2 - 12–14 years<br>3 - 15–17 years<br>4 - 18–20 years<br>5 - 21–24 years<br>6 - 25–29 years<br>7 - 30–95 years |
| GENDER | Gender (Male/Female) | 497 | 1 - Male<br>2 - Female |

Table 3. Description of dataset variables (Part c)

| Variable | Description | Missing Values | Encoded Value |
|---|---|---|---|
| LIVARAG | Living arrangements | 167,622 | 1 – Homeless<br>2 – Dependent Living<br>3 – Independent Living |
| LOS | Length of stay | 37 | 1-30 - 1-30 days (Different)<br>31 - 31 – 45 days<br>32 – 46 – 60 days<br>33 – 61 – 90 days<br>34 – 91 – 120 days<br>35 – 121 – 180 days<br>36 – 181 – 365 days<br>37 – More than 1 year |
| MARSTAT | Marital Status | 333,443 | 1 - Never married<br>2 - Now married<br>3 - Separated<br>4 - Divorced/Widowed |
| METHUSE | Medication assisted opioid therapy | 127,916 | 1 – Yes<br>2 - No |
| NOPRIOR | Previous treatment record | 148,309 | 0 – No prior treatment<br>1 – 1 or more treatments |
| PSOURCE | Primary source of referral | 81,287 | 1 - Individual<br>(includes self-referral)<br>2 - Alcohol/drug use<br>care provider<br>3 – Other health care provider<br>4 – School<br>5 – Employee<br>6 – Community referral<br>7 – Court/criminal/DWI/DUI |
| RACE | Race | 54,700 | 1 Alaska Native<br>2 American Indian<br>(other than Alaska Native)<br>3 Asian or Pacific<br>4 Black or African American<br>5 White<br>6 Asian<br>7 Other single races<br>8 Two or more races<br>9 Native Hawaiian<br>or another Pacific islander |

Table 4. Description of dataset variables (Part d)

| Variable | Description | Missing Values | Encoded Value |
|---|---|---|---|
| ROUTE1 | Route of administration (primary) | 111,263 | 1 – Oral<br>2 – Smoking<br>3 – Inhalation<br>4 – Injection<br>5 - Other |
| ROUTE2 | Route of administration (secondary) | 780,318 | 1 – Oral<br>2 – Smoking<br>3 – Inhalation<br>4 – Injection<br>5 - Other |
| ROUTE3 | Route of administration (tertiary) | 1,308,215 | 1 – Oral<br>2 – Smoking<br>3 – Inhalation<br>4 – Injection<br>5 - Other |
| SUB1 | Substance use at admission (primary) | 57,029 | 1 – None<br>2 – Alcohol<br>3 – Cocaine/Crack<br>4 – Marijuana<br>5 – Heroin<br>7 – Other opiates/synthetics<br>10 – Methamphetamine |
| SUB2 | Substance use at admission (secondary) | 64,788 | 1 – None<br>2 – Alcohol<br>3 – Cocaine/Crack<br>4 – Marijuana<br>5 – Heroin<br>7 – Other opiates/synthetics<br>10 – Methamphetamine |
| SUB3 | Substance use at admission (tertiary) | 150,301 | 1 – None<br>2 – Alcohol<br>3 – Cocaine/Crack<br>4 – Marijuana<br>5 – Heroin<br>7 – Other opiates/synthetics<br>10 – Methamphetamine |
| VET | Veteran Status | 204,589 | 1 – Yes<br>2 – No |
| REASON | Reason for discharge | None | 1 – Treatment Completed<br>2– Treatment Incomplete |

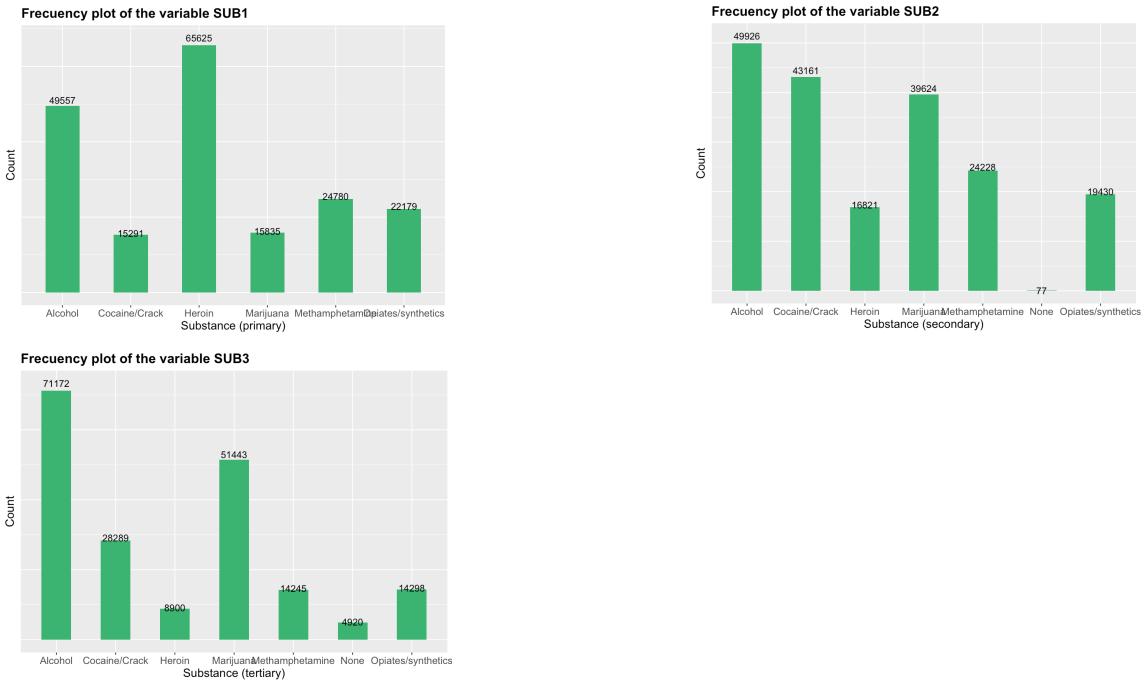


Figure 2. Plots displaying the Count of people and their drug preferences on a primary, secondary, and tertiary basis

The three images shown in Figure 2 are a comparison between the primary, secondary and tertiary substances. Heroin was the most common choice of consumption, followed by alcohol and marijuana. Analysis can be made wherein the primary choice is alcohol and then it is more of marijuana or heroin.

Figure 3 shows plots containing different types of substances consumed by the patients. Here, heroin is the most consumed primary substance, followed by alcohol and marijuana. In, the first plot, people consuming heroin do not complete their treatment as compared to patients consuming other drugs.

Figure 4 show the commonly type of treatments given to the patient. The non-intensive treatment is the most administered however, the success rate of completing this treatment is also low. Figure 5 indicates the length of days spent seeking treatment. Usually, patients tend to stay for 24 hours. Several factors come into picture and the treatment is independent to the length of stay at the facility as it depends on how well the patient is recovering.

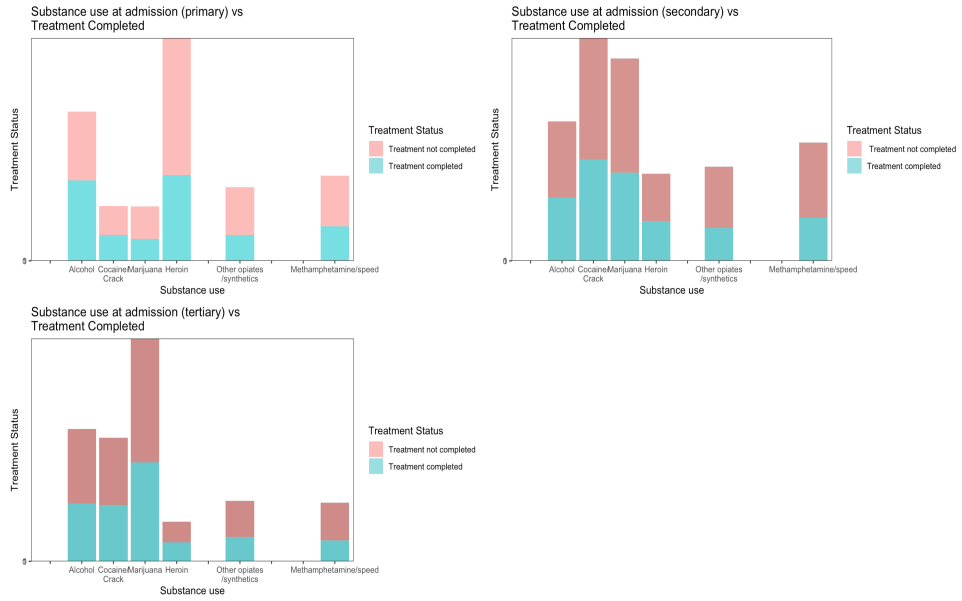


Figure 3. Plot displaying Substance in use vs Treatment Status, (0:Treatment Incomplete, 1:Treatment Complete)

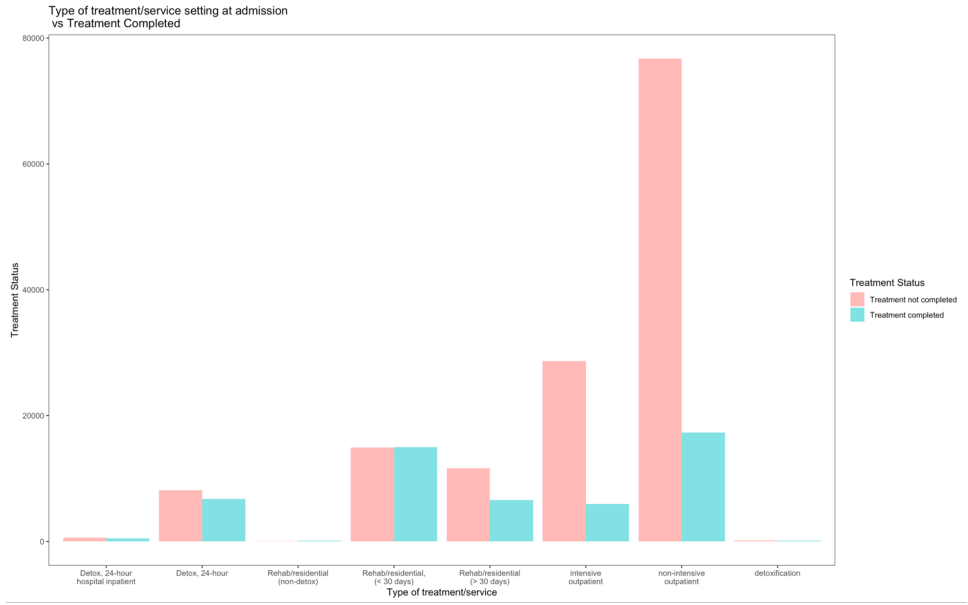


Figure 4. Plot displaying Type of treatment/service vs Treatment Status

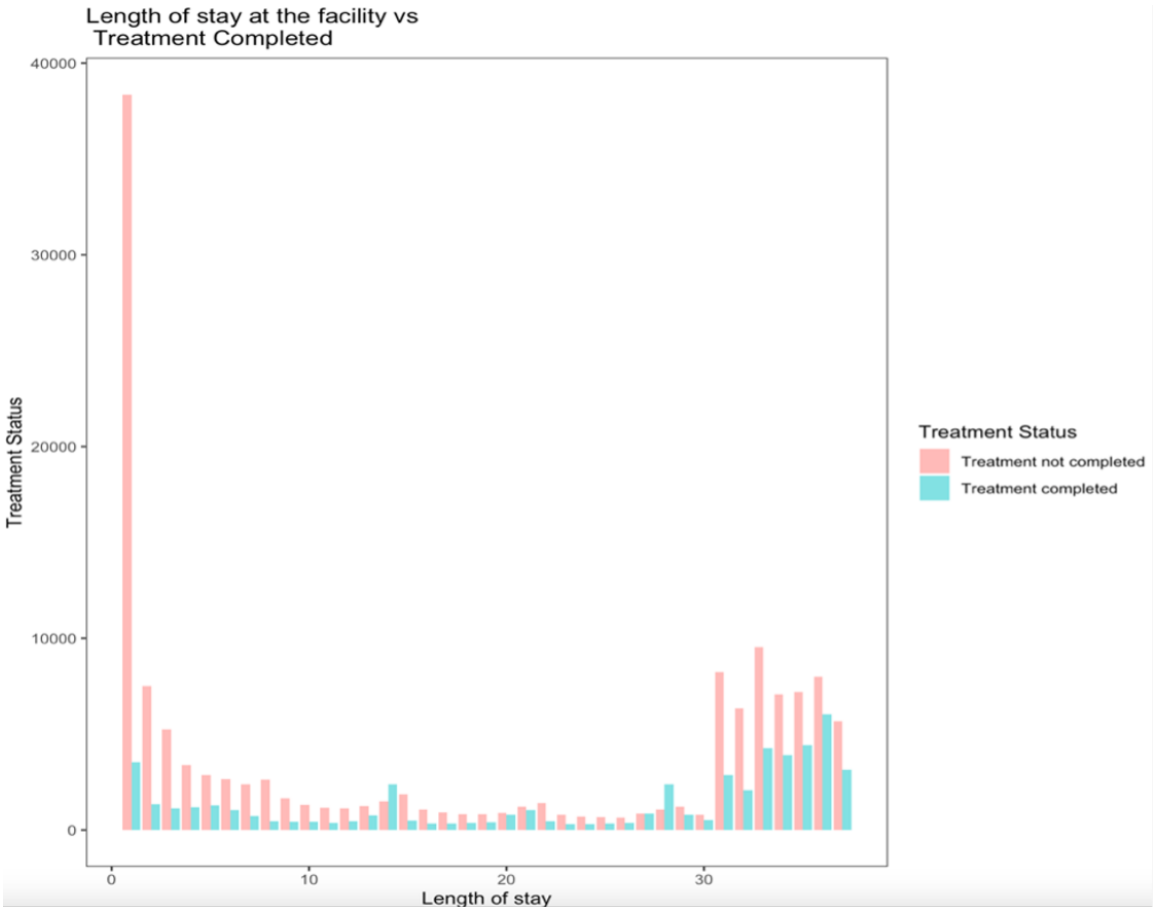


Figure 5. Plot showing Length of stay vs Treatment Status

## 4.2 Data pre-processing

### 4.2.1 Taking care of missing data

However, since the objective here is to create supervised machine learning models, which require values in the form of 0 and 1, all fields other than ”Treatment Completed” will be recorded as 0 indicating ”Treatment Incomplete”.

Missing data or null values can be handled in a variety of ways, including deleting records, imputing values using the mean or median, and imputing categorical values by replacing them with the most frequent value. In this dataset, all missing data values are represented as -9, and these values were handled by removing all rows from the entire dataset. Rows that had more than one important variable missing, such as substance use, age, frequency of use, etc., then the entire row is deleted Arndt, 2009. Columns such as AGE, SERVICES, ALCDRUG and CASEID did not contain any missing values. The disadvantage of imputing is that it does not take into account dependencies

between attribute values. Imputing a missing Length of stay (LOS) value with a commonly occurring value, for example, can deviate from its dependence on the type of service provided. In a scenario where missing LOS values are imputed with a value of 5 (which is more significant in services 1-3), it can have a significant impact on the outcome for services 4-8, where the patient taking more drugs would actually require more days to recover. The performance of the models was unaffected since the dataset's size was over 100,000 records. Using the dplyr package, 30 features were chosen with a total of 193,267 unique observations to be included in the analysis. The missing data is systematic as this data is not measured by the researcher. Some patients may intentionally leave out secondary and tertiary preferences of drugs consumed or the frequency of use.

### 4.2.2 Set the response variable as factors

The next step included in converting the response variable 'REASON' into factors. The response variable is a categorical variable that is either 0 or 1, and this is used to make predictions on the entire dataset. By converting it into factors using the as.factor(patient data$REASON) and storing it in the patient data$REASON, it is easy to visualize character vectors in a non-alphabetical order.

### 4.2.3 Feature selection

Feature selection is an important aspect in machine learning, and it is important to provide the right features while creating models to have better accuracy's and give better predictions. Without feature selection, the model's complexity is increased, and chances of over-fitting can occur. As a result, selecting only the relevant features that are non-redundant and uncorrelated can help the machine learning algorithm train faster, especially in cases such as the dataset used in this project must be trained.

### 4.2.4 Wrapper method using Boruta algorithm

The Boruta algorithm tries to capture features from the dataset that are important with respect to the response variable. This algorithm gives a call on the significance of features on the dataset. It tries to classify features as important and unimportant. Features that have not been assigned either of the classes are said to be tentative. For this dataset, all 29 features were deemed important, and the execution was completed after 12 iterations. Code listing 1 shows the count of important attributes in the dataset.

From the Figure 6 box plot the top five features are LOS (Length of stay), SERVICES, CASEID, DSCCRIT and PSOURCE.

```
> print(boruta)
Boruta performed 12 iterations in 46.76376 mins.
  29 attributes confirmed important: AGE, ALCDRUG, CASEID, DSMCRIT, EDUC and 24 more;
  No attributes deemed unimportant.
```

Listing 1. Code Output from the Boruta Algorithm

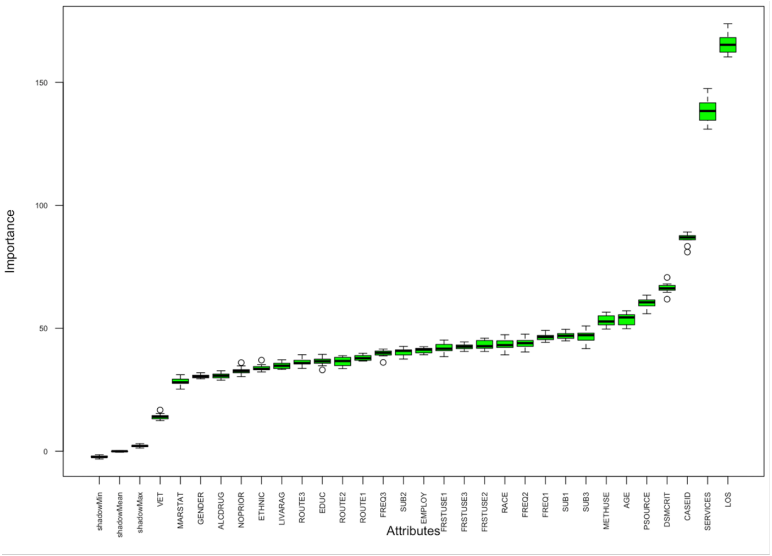


Figure 6. Box plot displaying Attributes ranked by Importance using the Boruta Algorithm

### 4.2.5 Using Random Forest

The Random Forest technique may be used to calculate the relative significance of each feature in a dataset that can be utilized for prediction. Random forest utilizes a tree-based technique that ranks nodes depending on how effectively they improve node purity. Nodes with the largest decrease in impurity appear at the beginning of the tree, whereas nodes with the least decrease appear at the end. A subset comprising the most important features can be formed by pruning beyond a specific node.

The top five features detected by the random forest method Figure 7 in this bar chart are the LOS, CASEID, SERVICES, AGE, and DSMCRIT. When both approaches' results were compared, the LOS CASEID and SERVICES were identified as the most essential. VET, ALCDRUG, and ETHIC were the lowest scored features.
In comparison to the Random forest, the Boruta algorithm wrapper method is more reliable. The results obtained from the Random Forest can vary depending on the randomness of the sample data that is used to create a model. In contrast to Random forest, the Boruta algorithm accurately shows which variables are important.

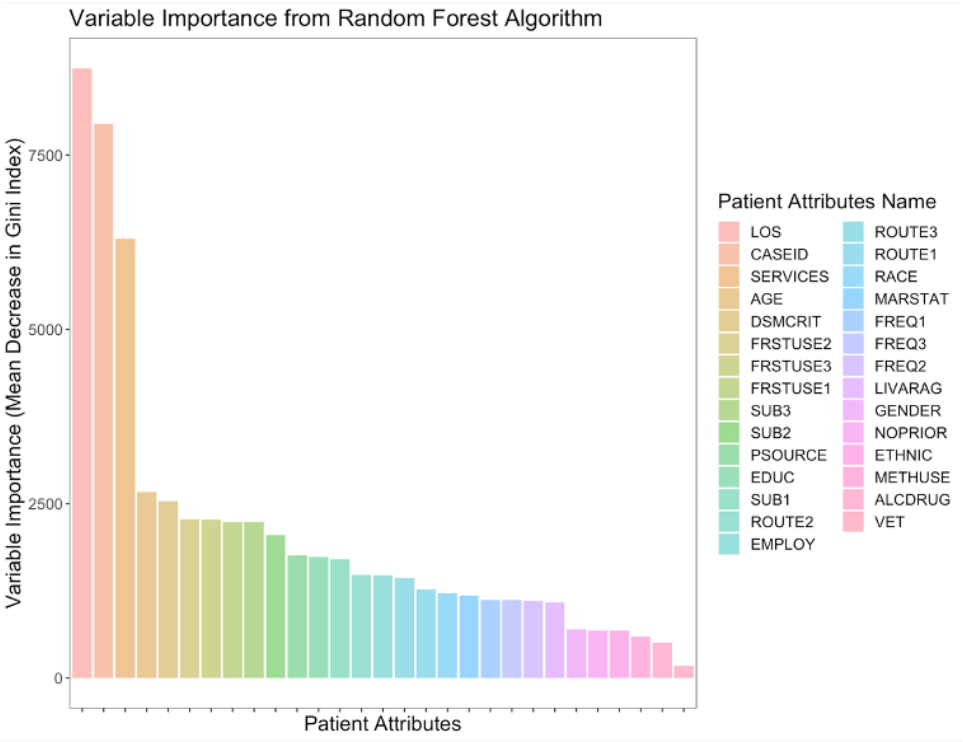


Figure 7. Bar chart displaying Attributes ranked by Importance using the Random Forest algorithm

### 4.2.6 Splitting data into train and test sets

The dataset is split into the train and test sets with the ratio of 80:20. A total of 154,614 observations were selected at random to the train set and 38,653 observations were given to the test set.

### 4.2.7 Feature scaling

Feature scaling was used to verify that the values are on the same scale or range and that no variable is dominated by another. All the variables were converted into features with values ranging from $-1$ to $+1$ using the scale() function. This was done primarily to prevent attributes like LOS from dominating other values due to their higher values.

### 4.2.8 Handling class imbalance

The dataset consists of imbalanced classes since the distribution is unequal and is highly skewed on the 0 (i.e., Treatment Incomplete). The reason behind this is due to converting all values to 0 under the REASON and preserving only the 1's (Treatment Complete. Upon viewing the confusion matrix for the Random Forest algorithm with the original train set, it was seen that the value for sensitivity (0.49) was extremely low with only 5176 successful treatment completed predictions whereas a high specificity for predicting 0's came to be around 26430 incomplete treatment predictions. The main aim of this project is to not predict 0's but predict 1's i.e., determine the maximum likelihood that a patient will complete the treatment and hence there is a need to have a balanced class to have better predictions. Hence, three different approaches were used to handle imbalanced dataset such as oversampling, under sampling and SMOTE techniques.

**i. Oversampling** Using the oversampling class distribution technique, it was possible to achieve a balanced dataset by randomly replicating instances from the majority class 0 onto the minority class 1. The library ROSE was used to implement this algorithm. A total of 225440 instances were replicated on the train and test set and distributed equally. There is no information loss however, the disadvantage of this method is that there is a chance for over-fitting. The accuracy recorded for this method 81.59 percent, sensitivity = 0.5645 and specificity = 0.9094.

**ii. Under sampling** The next method under sampling was implemented in which the observations are reduced from the larger class to create a balanced dataset. This method is useful when the dataset is huge and helps improve run time. The accuracy measured after implementing under sampling is 76.4 percent, sensitivity = 0.7910 and specificity = 0.7538. This method performed well in predicting the sensitivity i.e., 1's as compared to the original train set and over sampled dataset.

**iii. Synthetic minority sampling technique (SMOTE)**

Smote sampling is a technique used to prevent the model from over-fitting. A subset of data is taken from the minority class and with that newer instance are generated. These instances are then added to the original dataset. From this approach, the overall

accuracy came to be 78 percent, with sensitivity = 0.669 and specificity = 0.83. By comparing the three approaches, the SMOTE sampling termed to be the most appropriate choice to handle imbalanced dataset. Figure 8 depicts classes before and after dealing with class imbalance.

*Relation between accuracy, precision, recall and sensitivity:* Accuracy: is defined as the number of patients correctly labeled as Treatment Complete and Treatment Incomplete divided by the total number of patients correctly and incorrectly labeled as Treatment Complete and Incomplete. Precision: Number of patients correctly labeled as Treatment complete divided by the sum of the number of patients correctly and incorrectly labeled as treatment complete. Recall: Number of patients correctly labeled as Treatment complete divided by the sum of the number of patients correctly labeled as Treatment complete and incorrectly labeled as treatment incomplete. Specificity: Number of patients correctly labeled as Treatment incomplete divided by the sum of the number of patients correctly labeled as Treatment incomplete and incorrectly labeled as treatment complete

Given that this is a binary classification problem, measuring model accuracy is insufficient. It is also critical to distinguish between correct and incorrect classification. The goal is to accurately predict that a person who consumes more drugs requires more days to recover and is labeled as Treatment complete (True Negative) versus a person who consumes more drugs but has fewer treatment days and is labeled as Treatment incomplete (True positive). The specificity is preferred to cover true negative results for Treatment complete. With this information, the goal of this project is to determine what factors contribute to the result being classified as Treatment Complete.

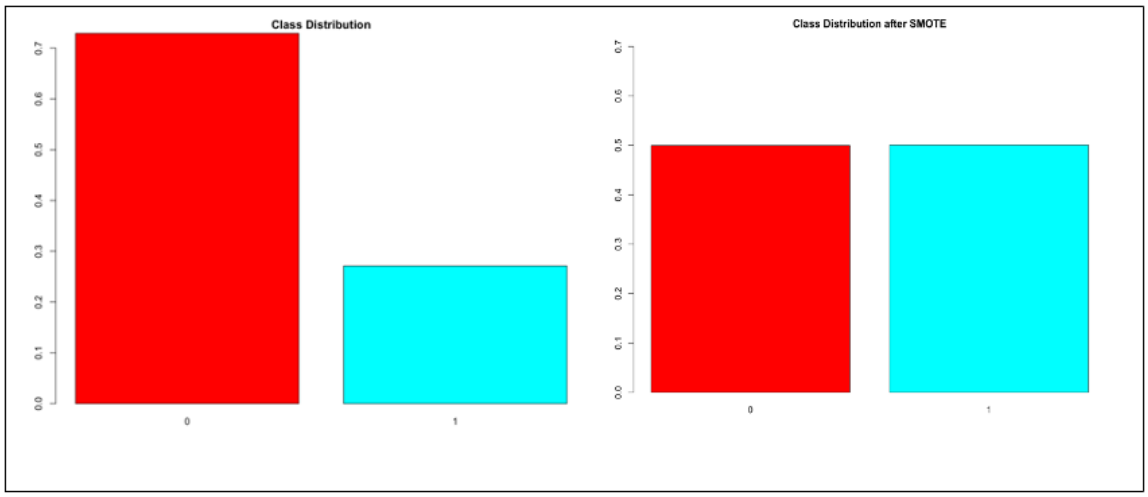


Figure 8. Bar chart displaying Class Distribution of Treatment Status before and after using the SMOTE technique

# 5 Machine Learning Algorithms

## 5.1 Logistic Regression

Logistic Regression is a supervised classification approach that assigns observations (features x) to one of a number of discrete classes (response y). Logistic regression is classified into two types: a. binary (0/1) and b. multi-linear functions (apple/ orange/ grapes). This is a predictive analytic method based on probability. Logistic regression is based upon the hypothesis that it tends to limit the cost function between 0 and 1 values. For any value beyond a certain threshold for example $> 0.5$ would mean that the function maps the value to 1 otherwise 0. This is an advantage since Linear functions tend to have probabilities $> 1$ or $< 0$. The Sigmoid function is used to map all the predictions to the probabilities.

For this project the Logistic Regression algorithm was implemented with the REASON as the target variable. All 29 features were considered while creating the model. This model was then evaluated with the test set to compute the accuracy of the model's performance. A threshold was set to be 0.5 for which any probabilities $> 0.5$ were indicated as class 1 otherwise class 0. The glm() function was used to implement the Logistic regression model and by specifying the parameter family = binomial.

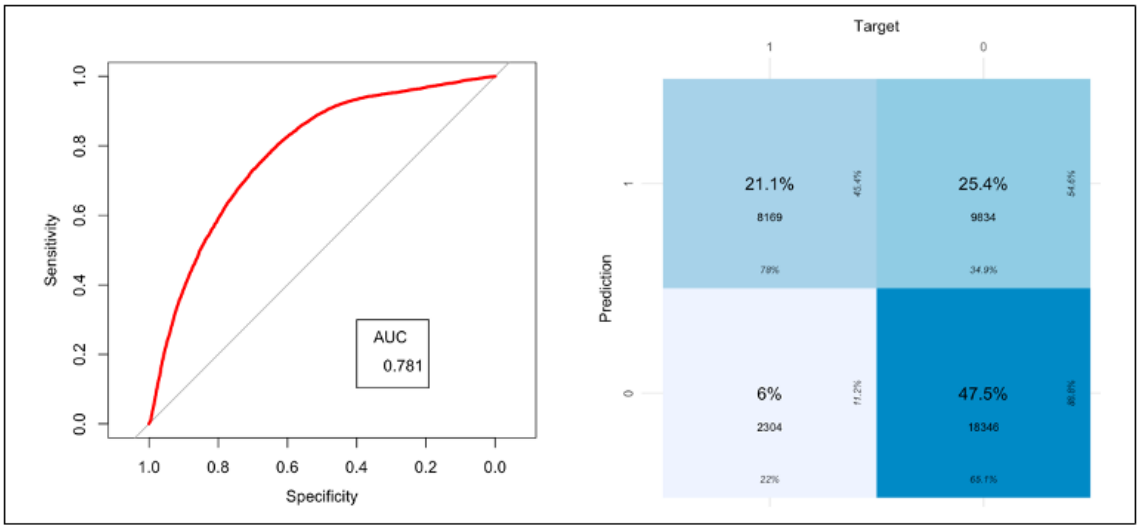


Figure 9. ROC Curve and Confusion Matrix for Logistic Regression synthetic dataset

Upon viewing the summary statistics for the synthetic dataset, features such as FIRSTUSE1, LIVARAG, FREQ3, ROUTE2 ROUTE3, SUB3 and VET have little significance as compared to the other features. View appendix A: A.1: Logistic Regression Accuracy for the confusion matrix code. According to the statistics in the Table 5, the original dataset has an test accuracy higher than the synthetic dataset. The sensitivity, i.e., predicting class 0 (treatment incomplete), was enhanced by utilizing synthetic

Table 5. Table comparing performance metrics of Original and Synthetic Train and Test Data sets using Logistic Regression

| | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|
| **Original Train Dataset** | 0.7820865 | 0.6135881 | 75.98 |
| **Original Test Dataset** | 0.7455261 | 0.5356544 | 75.950 |
| **Synthetic Train Dataset** | 0.8897604 | 0.4541429 | 68.62 |
| **Synthetic Test Dataset** | 0.8884262 | 0.4537577 | 68.59 |

data, but the specificity of predicting class 1 (Treatment complete) was only 0.45, leading us to conclude that Logistic Regression is not the appropriate model for making predictions for this dataset. The Receiver operating characteristics ROC curve for the synthetic dataset = 0.781 see Figure 9 for reference.

## 5.2 K – Nearest Neighbors

K-Nearest Neighbors (K-NN) is a supervised machine learning method that predicts output data points based on input data labels. The KNN method employs the notion of feature similarity, which it accomplishes by determining the proximity or similarity of the input data point to its neighbor and classifying it appropriately to the class that is highly similar in nature. Handling realistic data using KNN is more effective since it does not make assumptions about the given dataset (i.e., non- parametric).

K-Nearest Neighbors was created with the goal of classifying patients' treatment status as finished or incomplete. The 'class' package was installed to implement the KNN algorithm. The value of K was set to 393 i.e., taking the square root of the synthetic data train set. Other parameters were set to default. See appendix A, A.2: KNN synthetic data accuracy for the synthetic dataset implementation and to view the KNN implementation results.

Table 6. Table comparing performance metrics of Original and Synthetic Test and Train Data sets using K-Nearest Neighbors

| | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|
| **Original Train Dataset** | 0.7435 | 0.5256 | 75.22 |
| **Original Test Dataset** | 0.730111 | 0.486762 | 72.89 |
| **Synthetic Train Dataset** | 0.8224 | 0.5007 | 68.24 |
| **Synthetic Test Dataset** | 0.8438 | 0.4109 | 64.87 |

According to Table 6, the accuracy of the KNN model for the original train set was higher than that of the synthetic dataset. The Receiver Operating Characteristics ROC curve for the synthetic dataset is 0.66, as shown in Figure 10, which is lower than the

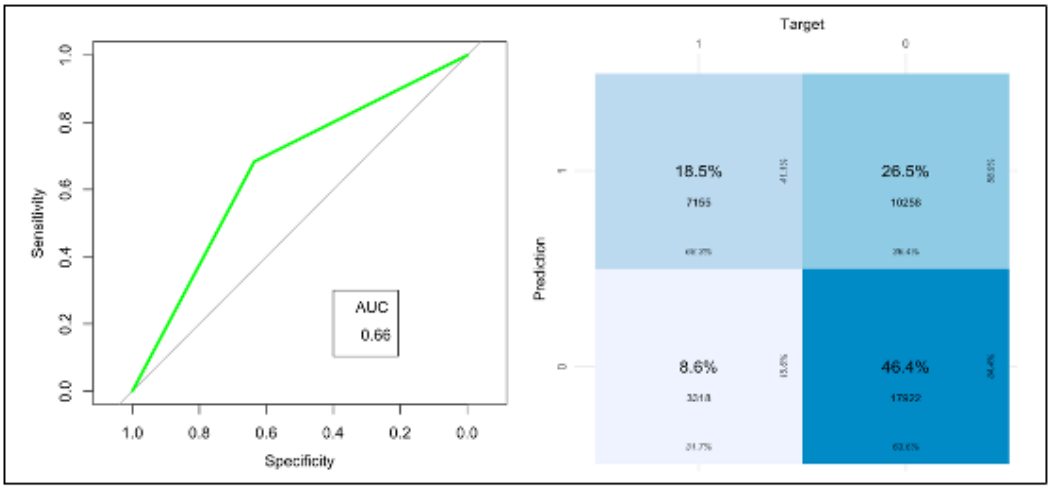


Figure 10. ROC Curve and Confusion Matrix for the KNN synthetic dataset

Logistic Regression model. Even after synthesizing new samples, the specificity for identifying class 1 is poor, while the sensitivity factor improves. When compared to the original training set, there is a greater loss in accuracy.

## 5.3 Naive Bayes

Naïve Bayes is also a supervised learning algorithm that consists of a family of simple probabilistic classifiers that are based on the Bayes theorem. It consists of strong ”nave” independent assumptions between the features of the dataset. The Naive Bayes method is named ”Naive” because it assumes that the occurrence of one characteristic is unrelated to the occurrence of other attributes. The idea behind the Bayes algorithm is that this theorem provides the conditional probability of an event A given the fact that another B event has already occurred. See Appendix A, A.3: Naive Bayes Accuracy code implementation results.

The aim here is to predict 0 (Treatment incomplete) or 1 (Treatment complete) depending on the set of 29 input features from the dataset. The package e1071 was loaded to implement the Naïve Bayes algorithm on and a predictive model was created on both the original and the synthetic dataset. Default values were chosen to implement this algorithm.

Bayes theorem works on the functionality:

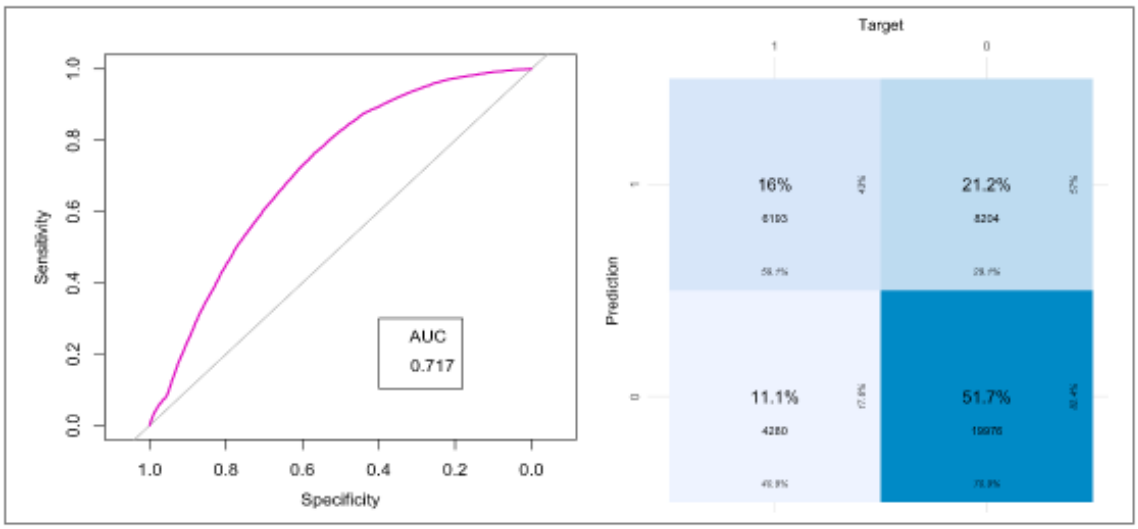


Figure 11. ROC Curve and Confusion Matrix for the Naive Bayes synthetic dataset

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \text{ where,}$$
$$P(A|B) = \text{Conditional probability of } A \text{ given } B$$
$$P(B|A) = \text{Conditional probability of } B \text{ given } A$$
$$P(A) = \text{Probability of event } A$$
$$P(B) = \text{Probability of event } B$$

According to Table 7, the accuracy of the original dataset was higher than that of the synthetic dataset. However, there seems to be little difference in the sensitivity and specificity for both data sets. The ROC curve has an AUC value of 0.717; see Figure 11 for ROC curve statistics. There appears to be some over-fitting between the original train and test dataset. Because the synthetic dataset has a low accuracy, this model will not be used for analysis.

Table 7. Table comparing performance metrics of Original and Synthetic Train and Test Data sets using Naive Bayes

| | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|
| **Original Train Dataset** | 0.7790545 | 0.4714878 | 71.76 |
| **Original Test Dataset** | 0.8235 | 0.486772 | 72.89 |
| **Synthetic Train Dataset** | 0.8235 | 0.429 | 67.63 |
| **Synthetic Test Dataset** | 0.8235154 | 0.4302 | 67.77 |

## 5.4 Decision Trees

The Decision Tree method is a supervised learning machine learning technique that may be used to solve classification and regression problems. This tree has the structure of an inverted tree, with each node representing a feature/predictor variable and the connections linking these nodes indicating decisions. Each leaf node represents the outcome, also known as the response variable. The advantage of employing a decision tree is that the findings are simple to interpret. Because decision trees employ a single feature per node to partition the data, they are faster to process. The library used to build the decision tree model is from the 'rpart' (Recursive Partitioning and Regression Trees) package and to build plots that 'rpart.plot' package and loaded onto the system. For this project, default values were used while implementing this algorithm on the original dataset and the method was set to 'class'. Predictions were made by the model and were compared with the test set to compute model accuracy using the confusion matrix.

The accuracy with the original train data turned out to be better as compared to the accuracy's with Logistic Regression, Naïve Bayes and K-NN. The decision tree has also done a better job in interpreting the class 1 values with a specificity = 0.6352. See Appendix A, A.4: Decision Tree original data accuracy code implementation.

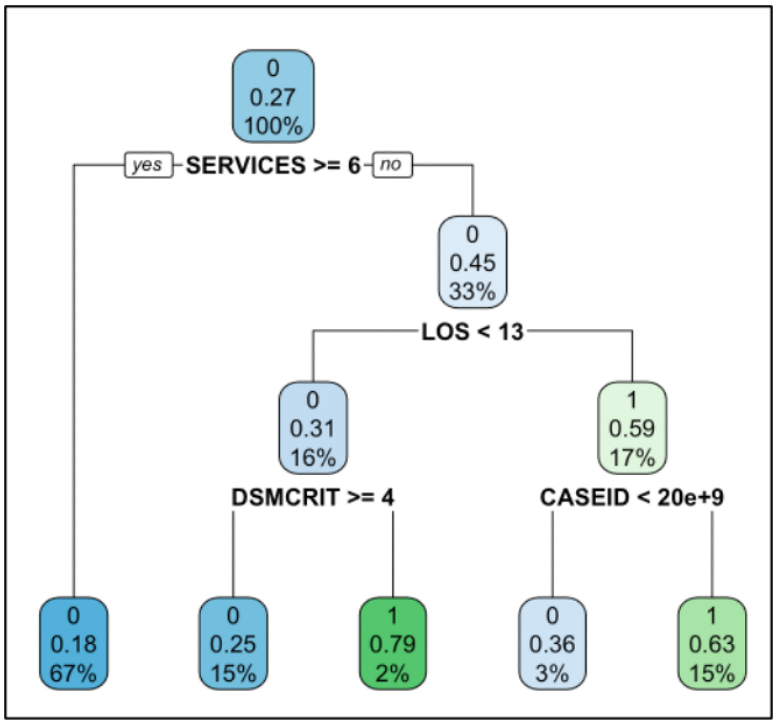


Figure 12. Plot for the Decision Tree algorithm on the original dataset

The decision trees in the Figure 12 were created with the rpart.plot() tool. We can derive from the above graph that if the services are >= 6 (6 = ambulatory, intensive

outpatient, 7 = ambulatory, non-intensive outpatient, or 8 = ambulatory, detoxification) and the LOS (Length of Stay in Days) is greater than 13 days at the facility and the DSMCRIT (Diagnostic and Statistical Manual of Mental Disorders diagnosis) is greater than 4, then any patient diagnosed with anything other than alcohol induced disorder/ substance use disorder/ alcohol dependence is more likely to complete their treatment which is represented as 1 otherwise they do not. Features like SERVICES, LOS, DSMCRIT, and CASEID were utilized to create this graph. However, it should be highlighted that this model is flawed because the case id is not crucial for efficient treatment completion.

Synthetic data is used to construct the decision tree to have more interpretations for a successful treatment. View Appendix A, A.5: Decision Tree accuracy synthetic code implementation. CASEID, LOS, METHUSE, PSOURCE, and SERVICES were obtained from the synthetic data to create the decision tree graph shown in Figure 13.

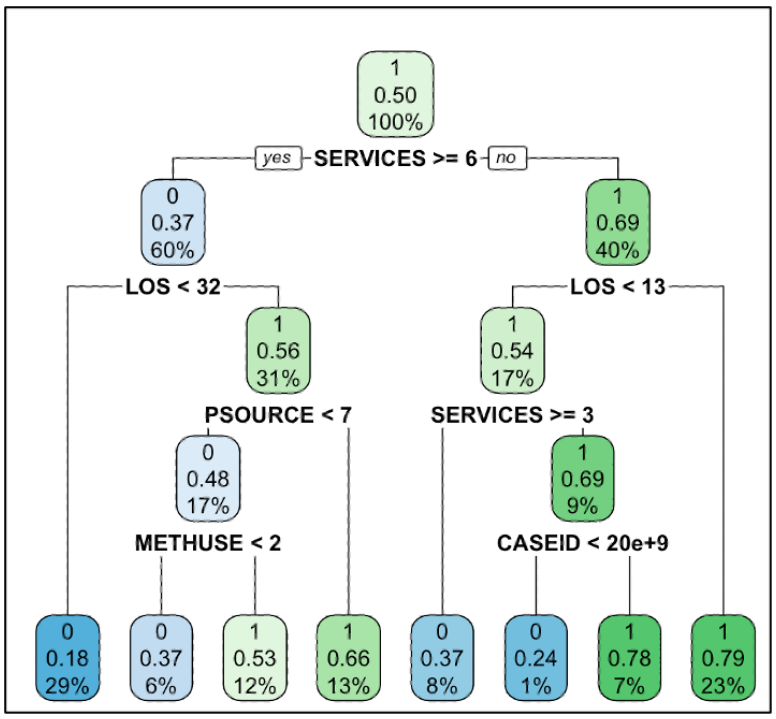


Figure 13. Plot for the Decision Tree algorithm on synthetic dataset

### 5.4.1 Pruning decision tree on the synthetic dataset

Pruning decision trees is a technique implemented in machine learning algorithms to reduce the size of the decision trees. This is done by pruning i.e., removing non-critical and redundant section of the trees which helps in improving the accuracy, reduces complexity and reduces over-fitting of noise from the train set. The cross-validation method

is implemented to determine the complexity parameter Cp. This is the amount through which splitting the node improved the relative error. This method is used as a stopping criterion and helps to prevent over-fitting. The code and graph values for varying Cp error rate values are shown in Figure 14. The code snippet shows the summary statistics and describes the cross validation error for each nsplit. This nsplit value is used to prune the tree. The lowest cross validation error denoted as xerror = 0.54358 with split = 7 is considered as the optimal value for Complexity parameter Cp = 0.01. View Appendix A, A.6: Decision Tree pruned accuracy code implementation.

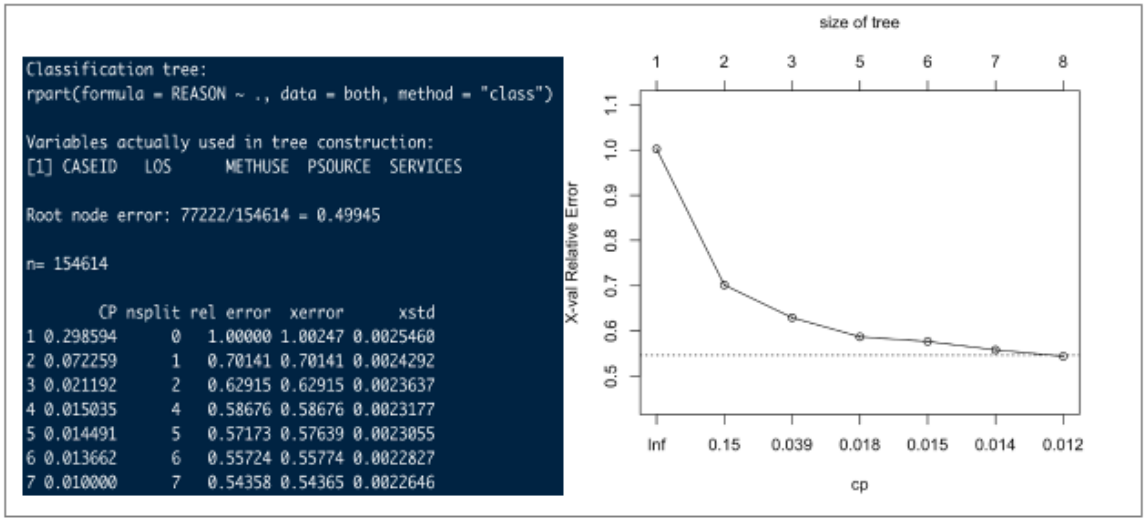


Figure 14. Number of variables used to construct the decision tree and Display Cp error rate

The plot shown in Figure 15 was created using the new Cp values. This Decision Tree suggests that patients with SERVICES >= 6 (6 = ambulatory, intensive outpatient, 7 = ambulatory, non-intensive outpatient, or 8 = ambulatory, detoxification) and LOS > =13 are more likely to complete their treatment otherwise only with the SERVICES >= then they would finish their treatment. If SERVICES such as (1 – Detox, 24-hour hospital inpatient, 2 – Detox, 24-hour hospital free standing residential, 3 – Rehab/residential (non-detox), 4 – Rehab/residential, ( 30 days), 5 – Rehab/residential, (> 30 days) along with LOS > 32 days with no prior referral are likely to finish treatment.

From the Figure 16 for the confusion matrix, the accuracy for the decision tree after pruning has increased by almost 4.69% (69.81% to 74.5%). By using cross validation approach and selecting the lowest error rate for Cp = 0.01, the accuracy of the model increased thereby lowering the error rate. Although the accuracy obtained is slightly lower than that obtained when training with the original train set, the pruned tree using synthetic data provides better predictions, as shown in Table 8.

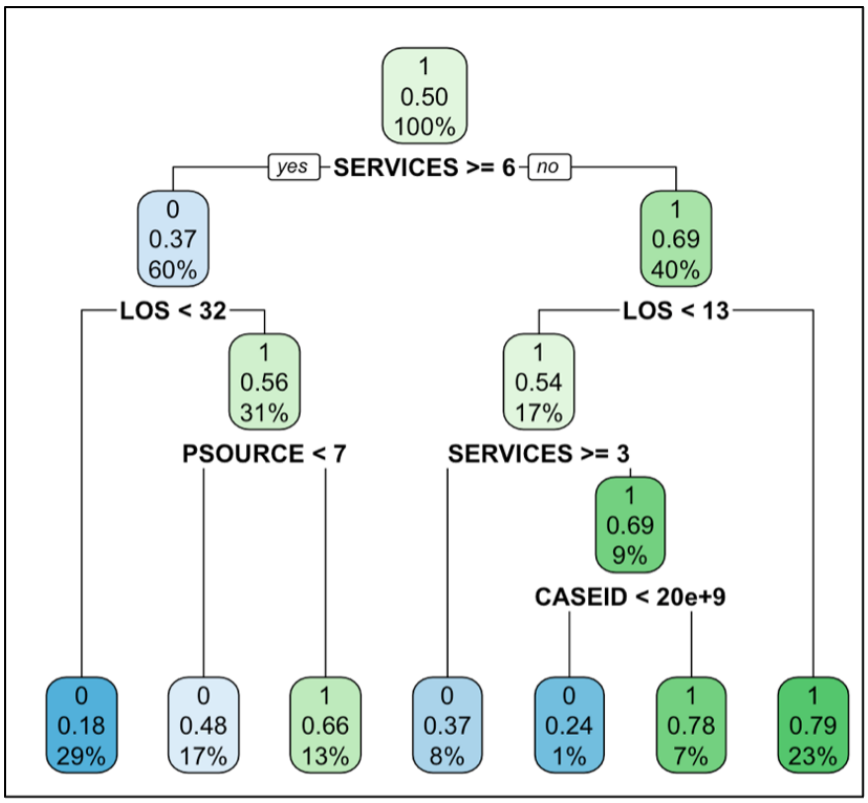


Figure 15. Pruned Decision Tree

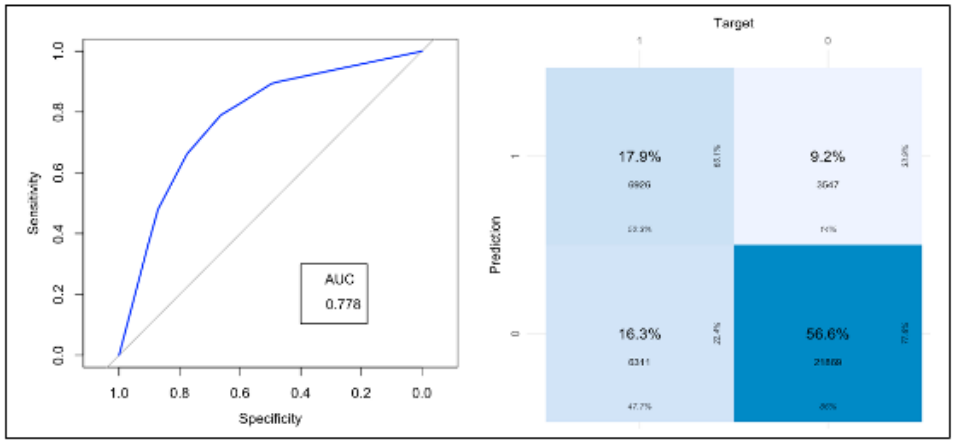


Figure 16. ROC Curve and Confusion Matrix for Pruned Decision Tree

Table 8. Table comparing performance metrics of Original and Synthetic Train and Test Data sets using Decision Tree

| | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|
| **Original Train Dataset** | 0.9208659 | 0.3864754 | 77.606 |
| **Original Test Dataset** | 0.7990 | 0.6352 | 77.26 |
| **Synthetic Train Dataset** | 0.6678141 | 0.7889674 | 70.06 |
| **Synthetic Test Dataset** | 0.89 | 0.4662 | 69.81 |
| **Pruned Train Dataset** | 0.7795688 | 0.6659665 | 74.87 |
| **Pruned Test Dataset** | 0.8604 | 0.5232 | 74.5 |

## 5.5 Random Forest (Core Algorithm)

Random Forest algorithms also known as bootstrap aggregation regression trees that can deal with models having high variance and low predictive powers. It is used for both classification and regression which creates a forest with many trees. The advantage of using Random Forest algorithm they have good performance and they do not require data pre – processing. They also have good OOB (Out of the bag) values. The only issue with this algorithm is that the execution time to train a model on a large data set is slow and can be less interpretable. This algorithm is preferred over decision trees since the accuracy is low. View Appendix A,A.7 to view the Random Forest code implementation.

This algorithm is evaluated on the original train dataset using over sampled, under sampled, and synthetic data. When implementing this algorithm, the default values were chosen. The Random Forest algorithm was used to find the maximum likelihood for predicting class 1's using the three approaches and have a balanced dataset.

Table 9. Table comparing performance metrics of Original and Synthetic Train and Test Data sets using Random Forest

| | **Sensitivity** | **Specificity** | **Accuracy** |
|---|---|---|---|
| **Original Train Dataset** | 0.9204 | 0.455 | 80.21 |
| **Original Test Dataset** | 0.4821 | 0.9117 | 79.53 |
| **Synthetic Train Dataset** | 0.7365 | 0.5228 | 78.24 |
| **Synthetic Test Dataset** | 0.6480 | 0.8087 | 76.50 |

View Appendix A, A.8 to view the code output after fine tuning the algorithm. The model using the original train set and the over sampled models were not included for further analysis due to over fitting and have poor predictions when introduced with new data. The under sampled data model had the least accuracy and hence was also not considered for further analysis. Hence, the synthetic data model appears to be the best choice in this scenario and hence will be considered for further tuning to make predictions. ROC curve statistics are shown in Figure 17.

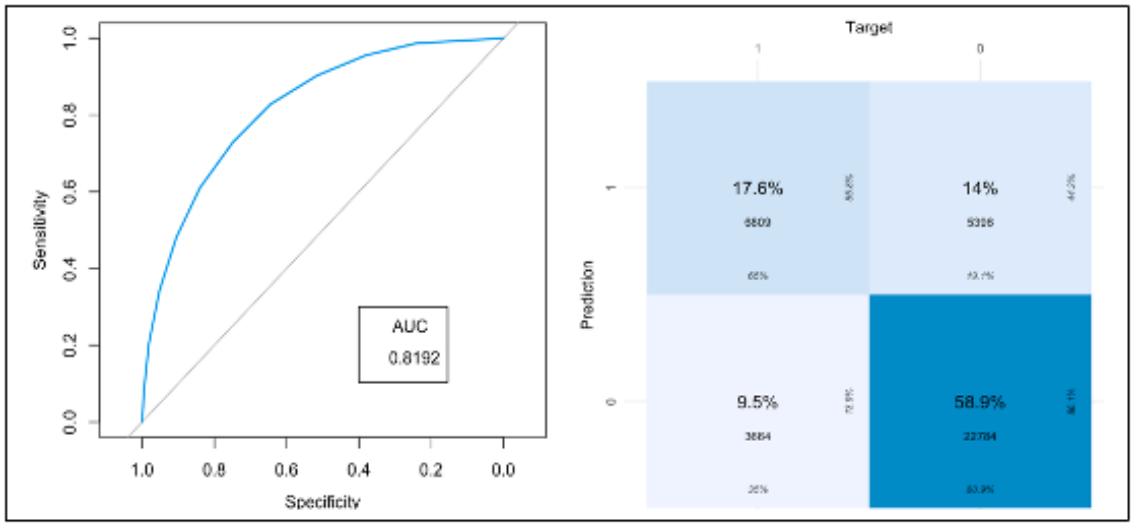


Figure 17. ROC Curve and Confusion Matrix for the Random Forest synthetic dataset

### 5.5.1 Fine tuning of Random Forest algorithm

Despite the fact that the pruned Decision tree outperformed the synthetic version of the Random Forest model in terms of accuracy, Decision trees are not always accurate. The plot for CASEID was shown in the previous tree in Figure 15, where CASEID is a unique identification number assigned to each patient. Making predictions based on a specific value in the CASEID feature is meaningless since it lacks critical information. Furthermore, Decision trees can be unstable especially when a small change is introduced, which in turn can have a major impact in obtaining the optimal decision tree. As a result, the Random Forest algorithm is implemented to generate more accurate predictions.

**Hyper-parameter tuning** The parameters such as ntree and mtry can be used to increase the model's predictive accuracy.

ntree: The ntree parameter indicates the number of trees to grow. Larger trees are complex and are computationally expensive to build. In this case, several values of ntree were testing to see if there was any change in accuracy. Upon comparing with the default value of ntree = 500, the accuracy of the model turned out to be 77

mtry: The mtry parameter indicates to the number of variables that need to be selected during a node split. To avoid the chances of over-fitting, ensure that the values of mtry are not small. The tuneRF function from the caret package is used to determine the recommended number of mtry values for the given dataset. The tuneRf method starts comparing mtry values at 3 and increases by a factor of 1.5 until the OOB error stops improving by 1%. The value of mtry is chosen based on the lowest OOB Error. The following code 2 shows the values for various mtry and its OOBError rate.

Upon comparing the plot from the Figure 18, and the code listing 2 we can conclude

```
> model_tuned <- tuneRF(
+   x=both[,-30],
+   y=both$REASON,
+   ntreeTry=500,
+   mtryStart=4,
+   stepFactor=1.5,
+   improve=0.01,
+   trace=FALSE
+ )
-0.04404461 0.01
0.03514197 0.01
0.01320901 0.01
0.002066929 0.01

> print(model_tuned)
       mtry   OOBError
3.OOB     3 0.07205686
4.OOB     4 0.06901704
6.OOB     6 0.06659164
9.OOB     9 0.06571203
13.OOB   13 0.06557621
```

Listing 2. Tune Random Forest algorithm to find the optimal mtry parameter

that the optimal value for mtry = 13 with the least OOB error rate = 0.06557621. Now implementing the same mtry = 13 on the Random Forest algorithm using the synthetic dataset we obtain the following results.

Based on these findings, we can conclude that the model's accuracy has been increased to 80%, with an estimated OOB error rate of 6.6 %, as shown in the code 3. The AUC values after tuning increased to 0.86, an improvement over the 0.81 shown in Figure 21. Figure 19 shows the plot where the error rate decreases as the size of the trees increases (ntree values). By implementing the varImp() we can identify the most important features that can be useful in obtaining a higher accuracy. Below is the plot Figure 20, displaying the top features selected by the algorithm. The LOS, SERVICES, FRSTUSE3, SUB3 and FRSTUSE2 are the top five features. By including features such as ALCDRUG, SUB1, GENDER, FREQ3 can lower the accuracy of the model. Similarly, the VET, ALCDRUG, ETHNIC can also decrease the Gini index values.

Therefore, by implementing parameter tuning, it was possible to achieve a higher accuracy by almost 3% on the synthetic version of data as shown in the Table 10. The ROC curve has an AUC value of 0.863 see Figure 21 which is improvement in the overall

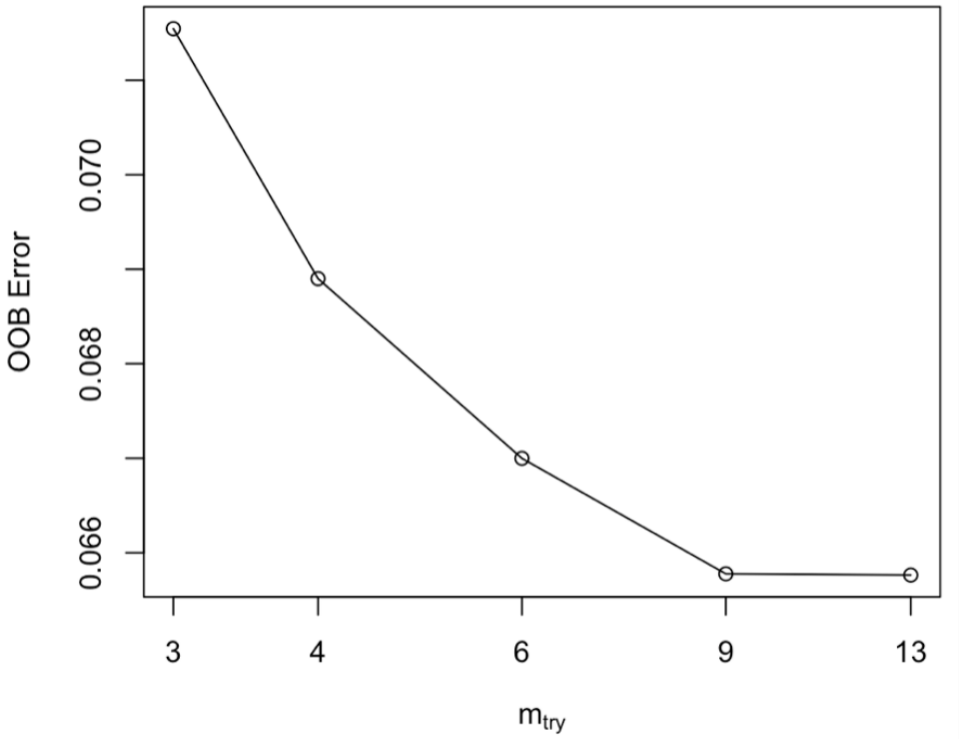


Figure 18. Plot showing the value for mtry vs OOB Error, (mtry: No of variables picked randomly at each split, OOB: Out of Bag Error)

model's performance.

```
> rf <-randomForest(REASON~.,data=both, mtry=13, importance=TRUE,ntree=500)
> print(rf)
Call:
 randomForest(formula = REASON ~ ., data = both, mtry = 13, importance = TRUE,  ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 13
        OOB estimate of  error rate: 6.6%
Confusion matrix:
      0     1 class.error
0 70727  6495  0.08410816
1  3707 73688  0.04786024
```

Listing 3. Accuracy after setting mtry = 13

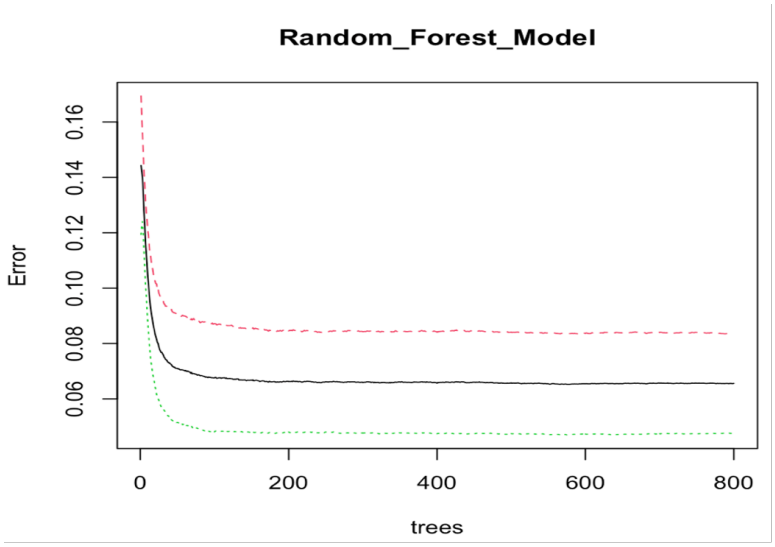


Figure 19. Plot displaying tree size vs error rate

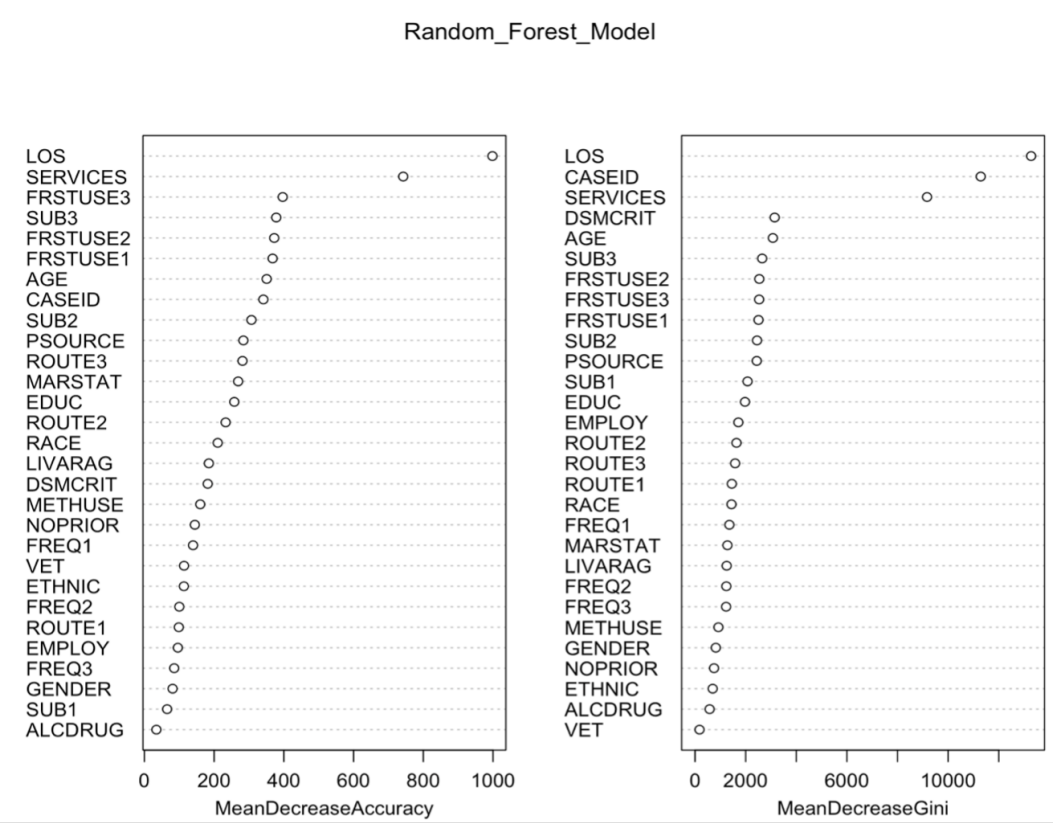


Figure 20. Plot displaying mean decrease of attributes using Accuracy and Gini Index

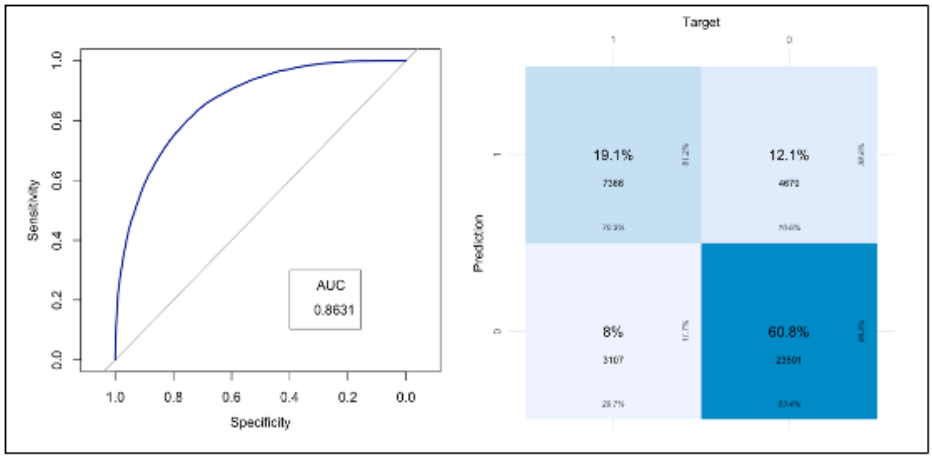


Figure 21. ROC Curve and Confusion Matrix for fine-tuned Random Forest dataset

Table 10. Table comparing performance metrics of both synthetic data before and after fine tuning

| | Sensitivity | Specificity | Accuracy |
|---|---|---|---|
| **Synthetic Train Dataset before fine tuning** | 0.9204 | 0.455 | 80.21 |
| **Synthetic Test Dataset before fine tuning** | 0.7580 | 0.6014 | 77.2% |
| **Synthetic Train Dataset after fine tuning** | 0.7994 | 0.6244 | 82.04% |
| **Synthetic Test Dataset after fine tuning** | 0.8339 | 0.7035 | 80% |

# 6 Results obtained using model predictions

According to the Random forest algorithm, the most important factor is the LOS (Length of stay) variable see Figure 20. Stacked bar charts are created to visualize how Length of stay relates to other variables such as the type of treatment and the type of substances consumed.

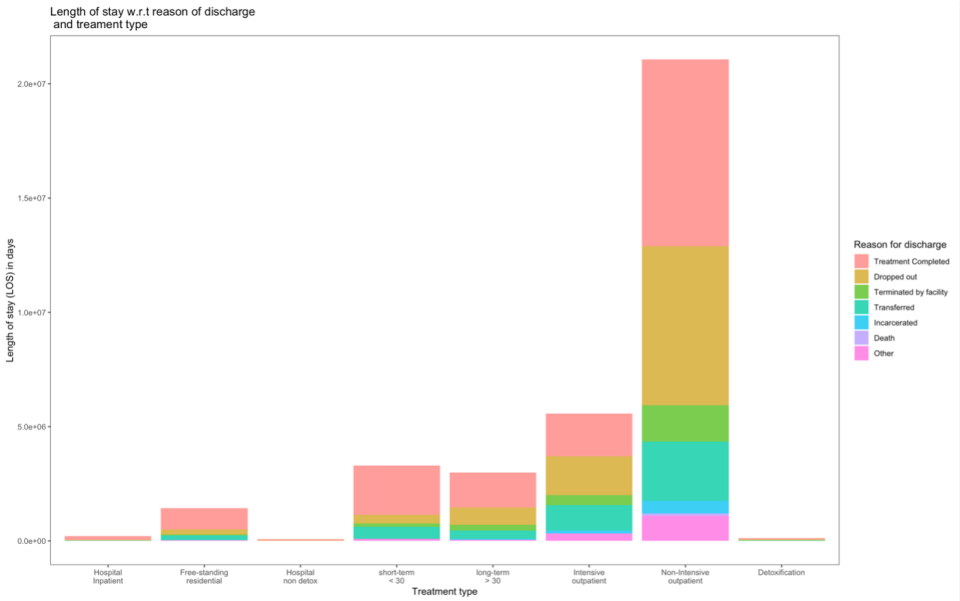


Figure 22. Bar chart indicating Length of stay w.r.t Reason of discharge and Treatment type

The bar plot in Figure 22 is plotted on the Treatment type vs Length of stay. Analysis can be made for a successful completion of a treatment the patient needs to stay for a longer period of time. Another bar graph is plotted on the Type of substances vs Length of stay see Figure 23. From this we can analyze that heroin and alcohol addictions are the most cases recorded amongst the other type of intoxicants consumed.

Various types of treatments are administered based on the level of consumption. The service types are numbered from 1-8 and for a detailed explanation view Appendix

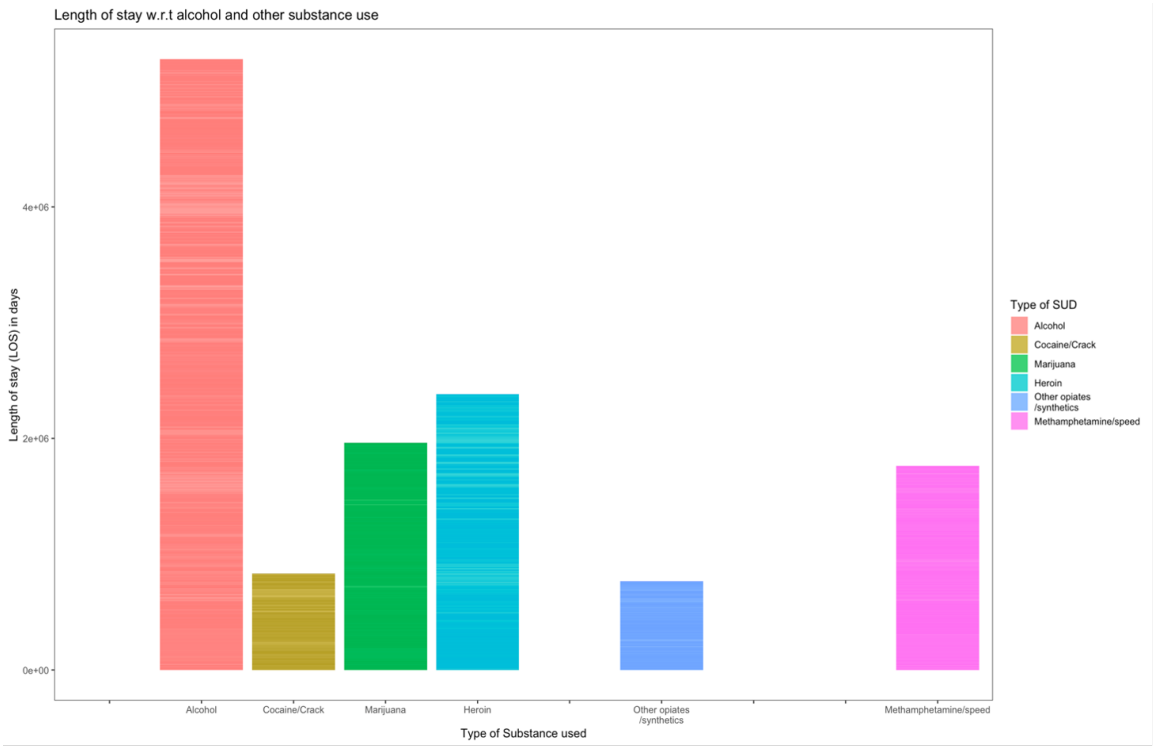


Figure 23. Bar chart indicating Substance in use vs Length of stay

B. Service 1 is provided to patients who have a low number of intoxicants as a result of drugs, whereas services 5-7 are provided to patients who have a high number of intoxicants.

Below are the table comparisons for each type of service with the type of addiction issues see Table 11, Table 12, Table 13, Table 14, Table 15, Table 16 and Table 17. Analysis was performed for each of the 7 types of alcohol/drug addiction by iterating through 8 different types of services and the length of stay factor. Other types of drugs were excluded from the analysis due to a lack of records in the dataset. By setting the frequency of use as 'daily use' and 'no use' in the new test dataset, predictions were made for single drug and multiple drug usage. This project aims to predict the number of days required for various types of addictions based on the type of treatment used.

Table 11. Table for Alcohol

| Alcohol | Service type | Length of stay 1 drug | Length of stay 2 or more drugs |
|---|---|---|---|
| | 1 | 3 | 5 |
| | 2 | 3 | 5 |
| | 3 | 6 | 16 |
| | 4 | 15 | 23 |
| | 5 | 19 | 27 |
| | 6 | 33 | 33 |
| | 7 | 33 | 33 |
| | 8 | 33 | 33 |

Table 12. Table for Cocaine/crack

| Cocaine/crack | Service type | Length of stay 1 drug | Length of stay 2 or more drugs |
|---|---|---|---|
| | 1 | 4 | 5 |
| | 2 | 4 | 6 |
| | 3 | 6 | 20 |
| | 4 | 16 | 23 |
| | 5 | 19 | 27 |
| | 6 | 33 | 33 |
| | 7 | 33 | 33 |
| | 8 | 33 | 33 |

Table 13. Table for Marijuana/hashish

| Marijuana | Service type | Length of stay 1 drug | Length of stay 2 or more drugs |
|---|---|---|---|
| | 1 | 5 | 5 |
| | 2 | 5 | 7 |
| | 3 | 7 | 16 |
| | 4 | 18 | 24 |
| | 5 | 22 | 27 |
| | 6 | 33 | 33 |
| | 7 | 33 | 33 |
| | 8 | 33 | 33 |

Table 14. Table for Heroin

| Heroin | Service type | Length of stay 1 drug | Length of stay 2 or more drugs |
|---|---|---|---|
| | 1 | 5 | 5 |
| | 2 | 6 | 6 |
| | 3 | 6 | 16 |
| | 4 | 18 | 24 |
| | 5 | 22 | 27 |
| | 6 | 33 | 33 |
| | 7 | 33 | 33 |
| | 8 | 33 | 33 |

Table 15. Table for Non-prescription methadone

| Non-prescription methadone | Service type | Length of stay 1 drug | Length of stay 2 or more drugs |
|---|---|---|---|
| | 1 | 5 | 5 |
| | 2 | 5 | 6 |
| | 3 | 6 | 16 |
| | 4 | 18 | 24 |
| | 5 | 22 | 27 |
| | 6 | 33 | 33 |
| | 7 | 33 | 33 |
| | 8 | 33 | 33 |

Table 16. Table for Other opiates and synthetics

| Other opiates and \synthetics | Service type | Length of stay 1 drug | Length of stay 2 or more drugs |
|---|---|---|---|
| | 1 | 5 | 14 |
| | 2 | 16 | 14 |
| | 3 | 15 | 16 |
| | 4 | 18 | 20 |
| | 5 | 20 | 27 |
| | 6 | 33 | 33 |
| | 7 | 33 | 33 |
| | 8 | 33 | 33 |

Table 17. Table for Methamphetamine/speed

| Methamphetamine /speed | Service type | Length of stay 1 drug | Length of stay 2 or more drugs |
|---|---|---|---|
| | 1 | 5 | 6 |
| | 2 | 7 | 7 |
| | 3 | 6 | 16 |
| | 4 | 19 | 20 |
| | 5 | 20 | 23 |
| | 6 | 33 | 33 |
| | 7 | 33 | 33 |
| | 8 | 33 | 33 |

The predictive model in Table 11, Table 12, Table 13, Table 14, Table 15, Table 16 and Table 17 displays the number of days required based on the service type for intoxicants such as Alcohol, Cocaine, Marijuana, Heroin, Non-prescription methadone, and other opiates. According to the model, different types of treatment have a different impact on the number of days required to recover from each type of addiction. The majority of drug addictions have similar recovery times, with a 2 to 3 day difference. According to these tables, the treatment cannot be successful if the patient does not stay for more than 10 days, especially if they are receiving Ambulatory, non-intensive outpatient (service 7) or Ambulatory, intensive outpatient (service 6). When patients receive these treatments, it indicates that the amount of drugs detected in their bloodstream is high, and these patients may require more time to recover, as indicated by the Length of stay for 2 or more drugs column. Similarly, when small amounts of cocaine or heroin are discovered, treatments like ambulatory detox take fewer days to recover.

The information displayed on the tables is not to compare drug types but to view the prediction of length of stay for all types of drugs based on the types of service provided. As a result, we can conclude that the minimum number of days specified in each of the tables for each service must be met in order for the treatment to be successful.

# 7 Conclusion

The goal of this project is to identify factors that can contribute to a patient's healthy recovery, depending on the services available. The primary goal of data pre-processing was to guarantee that the data from the US Department of Health and Human Services was in an acceptable format before we created our machine learning models. Preprocessing the data by splitting it into independent and dependent variables, dealing with missing data, analyzing essential features with the Boruta and Random Forest algorithms, and subsequently dealing with class imbalance using the SMOTE approach all led to the formation of reliable data to feed into our machine learning models. Exploratory data analysis was used to construct models, with the Random Forest model outperforming the others in terms of computing performance over large data sets, such as the 190,000+ rows of data in this project. When the performance metrics were compared, the Random Forest produced the best results in terms of accuracy, specificity, sensitivity, ROC, and AUC.The Random Forest algorithm, in comparison to other algorithms, was able to manage massive data sets due to characteristics such as OOB error detection and variable importance mechanisms. Hyper-parameter tuning further aided in improving the accuracy of the model to 80%. This model accurately predicts, up to 80% of the time, the minimum number of days required for a patient's successful recovery from drug addiction.

# 8 Limitations

One of the project's limitations is that the model has a 20% chance of being inaccurate when displaying the number of days required to complete the treatment. Using the SMOTE technique to generate synthetic data can result in class overlap, which can introduce additional noise as the number of overlapping classes increases. The Random Forest approach is computationally demanding in terms of both power and resources since it involves the construction of numerous trees to obtain an output. As it integrates decision trees to determine class, the training time increases with the number of trees.This project's limitation is the absence of quantitative data to establish the quantity of drugs used daily or weekly. With this information, it is possible to see variations in how long it takes for each drug to recover. Opiates, alcohol, and cocaine have the highest relapse rates, per the source Ciulla, 2021. According to this study, some drug types take longer to recover than others.

# 9 Future scope

The project's future goals include obtaining more valuable data including the patients history of consuming drugs and history of seeking care assistance to cure drug addiction. This data will help refine the analysis and help prescribe the best treatment to cure patients drug abuse addiction issues. The next goal is to improve the model's accuracy by fine-tuning it to more than 90%. To increase execution time, the number of estimators can be lowered by training the model with a larger value for trees and picking an optimum subset from it. Build a dashboard for quick and easy visual analysis.

# 10 Lessons Learned

Large dataset requires longer execution time and also requires more memory. Some of the possible solutions that can be done is using 1.Sampling only a subset of data, 2. Allocate more memory or use cloud services like AWS, 3.Making use of the RHadoop framework, which enables distributed processing of large data sets across computer clusters. Support Vector Machine algorithm works best for a smaller dataset as its training complexity is dependent on the size of the dataset. Hyper parameter tuning using Grid search on the entire dataset is computationally expensive. To test it out, select a random sample from the dataset and choose the top attributes. For faster data access and to prevent file corruption, large files should be stored in relational databases rather than in csv format. Before fitting a final model to the entire dataset, start with a smaller sample of it. Aids in quick turnaround of results and spot checks.

# References

Afzali, M. H., Sunderland, M., Stewart, S., Masse, B., Seguin, J., Newton, N., Teesson, M., & Conrod, P. (2019). Machine-learning prediction of adolescent alcohol use: A cross-study, cross-cultural validation. *Addiction*, *114*(4), 662–671.

Arndt, S. (2009). Stereotyping and the treatment of missing data for drug and alcohol clinical trials. *Substance Abuse Treatment, Prevention, and Policy*, *4*(1). https://doi.org/10.1186/1747-597x-4-2

Boslett, A. J., Denham, A., & Hill, E. L. (2020). Using contributing causes of death improves prediction of opioid involvement in unclassified drug overdoses in us death records. *Addiction*, *115*(7), 1308–1317.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique [Copyright - © 2002. Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the associated terms available at https://www.jair.org/index.php/-jair/about; Last updated - 2021-07-23]. *The Journal of Artificial Intelligence Research*, *16*, 321–357.

Ciulla, A. (2021). What drug has the highest relapse rate?: Beach house recovery. https://www.beachhouserehabcenter.com/what-drug-has-the-highest-relapse-rate/

Doyen, S., Taylor, H., Nicholas, P., Crawford, L., Young, I., & Sughrue, M. E. (2021). Hollow-tree super: A directional and scalable approach for feature importance in boosted tree models [Copyright - © 2021 Doyen et al. This is an open access article distributed under the terms of the Creative Commons Attribution License: http://creativecommons.org/licenses/by/4.0/ (the "License")]. *PLoS One*, *16*(10).

for disease control, C., & prevention 2022. (2022). *Opioid data analysis and resources @ONLINE*. https://www.cdc.gov/opioids/data/analysis-resources.html

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Juergens, J., & Hampton, D. (2021). Inpatient vs. outpatient rehab. https://www.addictioncenter.com/treatment/inpatient-outpatient-rehab/

Kamp, F., Proebstl, L., Hager, L., Schreiber, A., Riebschläger, M., Neumann, S., Straif, M., Schacht-Jablonowsky, M., Manz, K., Soyka, M., & Koller, G. (2019). Effectiveness of methamphetamine abuse treatment: Predictors of treatment completion and comparison of two residential treatment programs. *Drug and Alcohol Dependence*, *201*, 8–15. https://doi.org/https://doi.org/10.1016/j.drugalcdep.2019.04.010

Kelkar, K. M., & Bakal, J. W. (2020). Hyper parameter tuning of random forest algorithm for affective learning system. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 1192–1195. https://doi.org/10.1109/ICSSIT48917.2020.9214213

Kuhn, M., & Johnson, K. (2013). Over-fitting and model tuning. *Applied predictive modeling* (pp. 61–92). Springer.

Nath, P., Kilam, S., & Swetapadma, A. (2017). A machine learning approach to predict volatile substance abuse for drug risk analysis. *2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 255–258.

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC bioinformatics*, *11*(1), 1–13.

on Drug Abuse, N. I. (2020). Treatment and recovery. https://www.drugabuse.gov/publications/drugs-brains-behavior-science-addiction/treatment-recovery

Pitchers, K. K., Sarter, M., & Robinson, T. E. (2018). The hot 'n'cold of cue-induced drug relapse. *Learning & Memory*, *25*(9), 474–480.

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(3), e1301.

SAMHDA. (2021). *Samhda,substance abuse and mental health data archive @ONLINE*. https://www.datafiles.samhsa.gov/dataset/teds-d-2019-ds0001-teds-d-2019-ds0001

Tapia-Galisteo, J., Iniesta, J. M., Pérez-Gandía, C., García-Sáez, G., Puértolas, D. U., Izquierdo, F. J., & Hernando, M. E. (2020). Prediction of cocaine inpatient treatment success using machine learning on high-dimensional heterogeneous data. *IEEE Access*, *8*, 218936–218953. https://doi.org/10.1109/ACCESS.2020.3041895

Thompson, H. M., Faig, W., VanKim, N. A., Sharma, B., Afshar, M., & Karnik, N. S. (2020). Differences in length of stay and discharge destination among patients with substance use disorders: The effect of substance use intervention team (suit) consultation service [Name - Rush University Medical Center; Community Behavioral Health; Copyright - © 2020 Thompson et al. This is an open access article distributed under the terms of the Creative Commons Attribution License: http://creativecommons.org/licenses/by/4.0/ (the "License")]. *PLoS One*, *15*(10).

Wang, Q. Q., Kaelber, D. C., Xu, R., & Volkow, N. D. (2021). Covid-19 risk and outcomes in patients with substance use disorders: Analyses from electronic health records in the united states. *Molecular psychiatry*, *26*(1), 30–39.

Wetherill, R. R., Rao, H., Hager, N., Wang, J., Franklin, T. R., & Fan, Y. (2019). Classifying and characterizing nicotine use disorder with high accuracy using machine learning and resting-state fmri. *Addiction biology*, *24*(4), 811–821.

Zarkin, G. A., Dunlap, L. J., Bray, J. W., & Wechsberg, W. M. (2002). The effect of treatment completion and length of stay on employment and crime in outpatient drug-free treatment. *Journal of Substance Abuse Treatment*, *23*(4), 261–271.

# A Code listings

## A.1 Logistic regression accuracy

```
$ prob_predict2 <- predict(lr_both,
   type = 'response', newdata = test_set[-30])
> y_pred2 <- ifelse(prob_predict2 > 0.5, 1, 0)
> print(cm <- table(test_set[, 30], y_pred2))
   y_pred2
        0     1
  0 18346  9834
  1  2304  8169
> print("Logistic Regression")
[1] "Logistic Regression"
> print(paste("Accuracy of the test set: ",
(sum(diag(cm))/sum(cm))*100, "%"))
[1] "Accuracy of the test set:  68.5975215377849 %"
> print(paste("Error rate of the test set: ",
(1-sum(diag(cm))/sum(cm))* 100, "%"))
[1] "Error rate of the test set:  31.4024784622151 %"
```

## A.2 KNN Confusion Matrix synthetic

```
confusionMatrix(table(test_set[, 30], knn_both))
Confusion Matrix and Statistics

   knn_both
        0     1
  0 17922 10258
  1  3318  7155

               Accuracy : 0.6488
                 95% CI : (0.644, 0.6535)
    No Information Rate : 0.5495
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2642

 Mcnemar's Test P-Value : < 2.2e-16
```

```
            Sensitivity : 0.8438
            Specificity : 0.4109
         Pos Pred Value : 0.6360
         Neg Pred Value : 0.6832
             Prevalence : 0.5495
         Detection Rate : 0.4637
   Detection Prevalence : 0.7291
      Balanced Accuracy : 0.6273

       'Positive' Class : 0
```

## A.3 Naive Bayes

```
confusionMatrix(table(test_set[, 30], naive_pred1))
Confusion Matrix and Statistics

   naive_pred1

       0     1
  0 19976  8204
  1  4280  6193

               Accuracy : 0.677
                 95% CI : (0.6723, 0.6817)
    No Information Rate : 0.6275
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2686

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8235
            Specificity : 0.4302
         Pos Pred Value : 0.7089
         Neg Pred Value : 0.5913
             Prevalence : 0.6275
         Detection Rate : 0.5168
```

```
   Detection Prevalence : 0.7291
      Balanced Accuracy : 0.6269

       'Positive' Class : 0
```

## A.4 Decision Tree

```
Confusion Matrix and Statistics

   predictions0
        0     1
  0 25910  2270
  1  6520  3953

               Accuracy : 0.7726
                 95% CI : (0.7684, 0.7768)
    No Information Rate : 0.839
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3403

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.7990
            Specificity : 0.6352
         Pos Pred Value : 0.9194
         Neg Pred Value : 0.3774
             Prevalence : 0.8390
         Detection Rate : 0.6703
   Detection Prevalence : 0.7291
      Balanced Accuracy : 0.7171

       'Positive' Class : 0
```

## A.5 Decision tree Accuracy Synthetic

```
ConfusionMatrix(table(test_set[, 30], predictions1))
Confusion Matrix and Statistics
```

```
  predictions1
       0     1
0 18708  9472
1  2199  8274

               Accuracy : 0.6981
                 95% CI : (0.6935, 0.7026)
    No Information Rate : 0.5409
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3726

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8948
            Specificity : 0.4662
         Pos Pred Value : 0.6639
         Neg Pred Value : 0.7900
             Prevalence : 0.5409
         Detection Rate : 0.4840
   Detection Prevalence : 0.7291
      Balanced Accuracy : 0.6805

       'Positive' Class : 0
```

## A.6 Pruned decision tree accuracy

```
Confusion Matrix and Statistics

  dt_pred
       0     1
0 21869  6311
1  3547  6926

               Accuracy : 0.745
                 95% CI : (0.7406, 0.7493)
    No Information Rate : 0.6575
    P-Value [Acc > NIR] : < 2.2e-16
```

```
                 Kappa : 0.4039

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8604
            Specificity : 0.5232
         Pos Pred Value : 0.7760
         Neg Pred Value : 0.6613
             Prevalence : 0.6575
         Detection Rate : 0.5658
   Detection Prevalence : 0.7291
      Balanced Accuracy : 0.6918

       'Positive' Class : 0
```

## A.7 Random Forest

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 23425  3116
         1  4755  7357

               Accuracy : 0.7964
                 95% CI : (0.7923, 0.8004)
    No Information Rate : 0.7291
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5087

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7025
            Specificity : 0.8313
         Pos Pred Value : 0.6074
         Neg Pred Value : 0.8826
```

```
            Prevalence : 0.2709
        Detection Rate : 0.1903
  Detection Prevalence : 0.3134
     Balanced Accuracy : 0.7669

      'Positive' Class : 1
```

## A.8 Random Forest algorithm after fine tuning

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 23532  3109
         1  4648  7364

               Accuracy : 0.7993
                 95% CI : (0.7953, 0.8033)
    No Information Rate : 0.7291
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5145

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7031
            Specificity : 0.8351
         Pos Pred Value : 0.6131
         Neg Pred Value : 0.8833
             Prevalence : 0.2709
         Detection Rate : 0.1905
   Detection Prevalence : 0.3108
      Balanced Accuracy : 0.7691

      'Positive' Class : 1
```

# B Explanation of each type of service

## B.1 Service 1 : Detox, 24-hour, hospital inpatient

This type of treatment consists of providing Medical emergency care services in a hospital environment that is available 24 hours a day for detoxification of those who have significant medical issues from withdrawal.

## B.2 Service 2: Detox, 24-hour, hospital outpatient

This type of treatment is also given for 24 hours of the day which takes place in a non hospital setting that allows the patients to have safer withdrawal and also transition them to further treatments.

## B.3 Service 3: Rehab/residential, hospital (non-detox)

The treatment is given for 24 hours of medical care that takes place inside the hospital which also provides treatment for alcohol and other dependencies.

## B.4 Service 4: Rehab/residential, short term

This treatment is given to patients that consists of providing non acute care for alcohol and other dependencies.

## B.5 Service 5: Rehab/residential, long term

Longer period of time is given to this type of treatment which consists of non acute type of services. This can include living in places such as a halfway houses.

## B.6 Service 6: Ambulatory, intensive outpatient

This type of treatment consists of providing treatments and help with withdrawal that takes place in a hospital setting.

## B.7 Service 7: Ambulatory,non-intensive outpatient

This type of treatment service includes pharmacological therapies. It involves the individual along with families or group sessions.

## B.8 Service 8: Ambulatory, detoxification

This treatment involves in an outpatient setting consisting of both pharmacological and non-pharmacological settings.