Contents

[1 OBJECTIVE 2](#_Toc135608945)

[1.1 Predictor variable 2](#_Toc135608946)

[1.2 Model Details 2](#_Toc135608947)

[1.3 METHODOLOGY 2](#_Toc135608948)

[2 EXPLORATORY DATA ANALYSIS 3](#_Toc135608949)

[2.1 Univariate Analysis 3](#_Toc135608950)

[2.2 Population success vs Region Success 7](#_Toc135608951)

[2.3 Successfully treated drug 8](#_Toc135608952)

[2.4 State wise variation. 10](#_Toc135608953)

[3 MODEL 12](#_Toc135608954)

[4 CONCLUSION AND RECOMENDATION 12](#_Toc135608955)

[4.1 Recommendation Regarding Assumption 12](#_Toc135608956)

[4.2 Model Improvement 12](#_Toc135608957)

List of Figures

[Figure 2‑1 Education on Success Rate 3](#_Toc135608906)

[Figure 2‑2 Education on Success Rate 4](#_Toc135608907)

[Figure 2‑3 Success Rate based on primary Substance 5](#_Toc135608908)

[Figure 2‑4 Success Rate based on Secondary Substance 6](#_Toc135608909)

[Figure 2‑5 Region on Success rate 7](#_Toc135608910)

[Figure 2‑6 Division on Success rate 8](#_Toc135608911)

[Figure 2‑7 Success probability for diff drugs 9](#_Toc135608912)

[Figure 2‑8 Average time for the success of diff drugs 10](#_Toc135608913)

[Figure 2‑9 Ratio of population to rehab centers. 11](#_Toc135608914)

# OBJECTIVE

Evaluate substance abuse treatment across the entire United States. It may be used for evaluating the drug treatment on a person before whether the program is going to be a success or not. Apart from this we had a few questions that needed to be answered with the provided data.

* 1. What is the difference in treatment completion rates between different types of substances?
  2. Which substances have the greatest differences in the probability of treatment completion and the shortest completion times?
  3. Is this consistent across states/statistical areas or are there states/statistical areas that have policies that may suggest why these trends are occurring?

## Predictor variable

It started with finding out the predictor variable. And the predictor variable here is **REASON.** It has 7 different categories shown below.

1. Treatment completed: Successfully Treated
2. Dropped out of treatment: Unsuccessful.
3. Terminated by facility: Unsuccessful.
4. Transferred to another treatment program or facility: Confusing but will consider it as successful treatment as it cleared current treatment program.
5. Incarcerated: Unsuccessful
6. Death: Unsuccessful
7. Other: Unsuccessful

## Model Details

So, we performed binary classification with logistic regression and gradient boosting algorithm and considered accuracy to the deciding metrics. It was found that Logistic regression is quite accurate with an accuracy of 70% while ensemble algorithm i.e., gradient boosting had an accuracy of 63%.

Considering time and accuracy LR is significantly good model.

## METHODOLOGY

The data provided and meta data was clean and meta data which made it easier to understand. After understanding the data, we were able to explore it and succeeded in drawing reasonable insight. From these insights we were able to create a classifier model. We started with the simplest model i.e., Logistic regression and got good accuracy with this.

# EXPLORATORY DATA ANALYSIS

## Univariate Analysis

To gain some insights about demography, geography etc. data was plotted between these variables. Some of the most critical features are plotted below.

A picture containing text, diagram, screenshot, plot

Description automatically generated

Figure 2‑1 Education on Success Rate

Figure 2‑1 shows the dependency of education on treatment rate and it shows that the rate increases as the education increases. Similarly, all the variables were plotted.

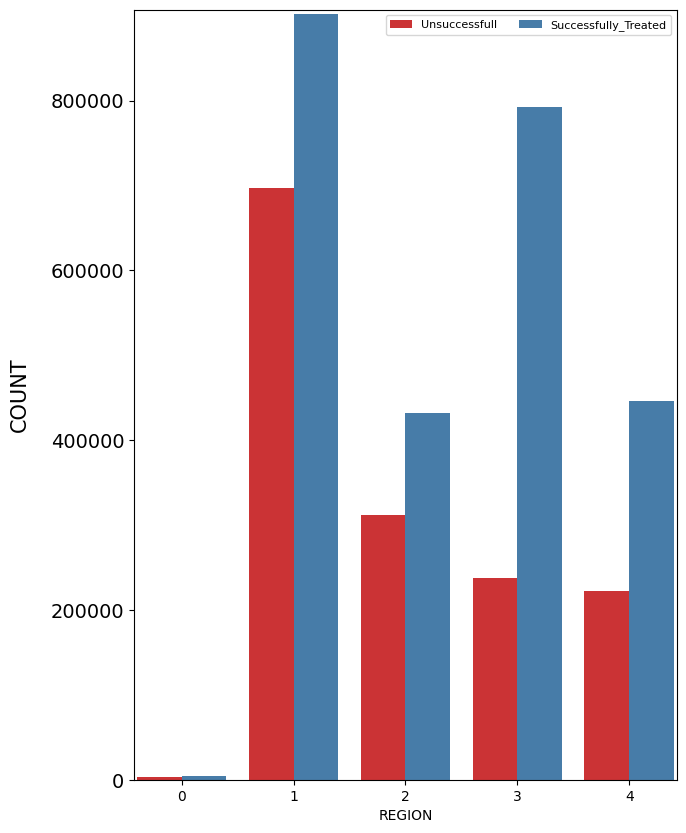


Figure 2‑2 Education on Success Rate

Figure 2‑2 shows the treatment rate across 4 regions and Southern territories shows. Higher success rate.

A picture containing screenshot, plot, diagram, text

Description automatically generated

Figure 2‑3 Success Rate based on primary Substance

A picture containing screenshot, plot, line, diagram

Description automatically generated

Figure 2‑4 Success Rate based on Secondary Substance

Figure 2‑3 and Figure 2‑4 shows the treatment rate based on substance former is based on primary substance while later is secondary substance. It can be concluded that success rate for alcohol and heroine is good as compared to other substances.

## Population success vs Region Success

Relationships were established based on proportionality as well, so we checked the success rate based on different factors to the population proportion.

For example, if success rate for males is 60% and overall success is also about 63% so there is not much change so the effect of these variable would not be that significant. We concentrated on regions and found different states that are not doing well.

A picture containing text, screenshot, number, font

Description automatically generated

Figure 2‑5 Region on Success rate

Figure 2‑5 shows that the success rate in southern region is more than northeastern region. So, the govt. is doing good there to achieve this.

Figure 2‑6 shows that the success rate in East south-central region is more than middle Atlantic region. So, the govt. is doing good there to achieve this.

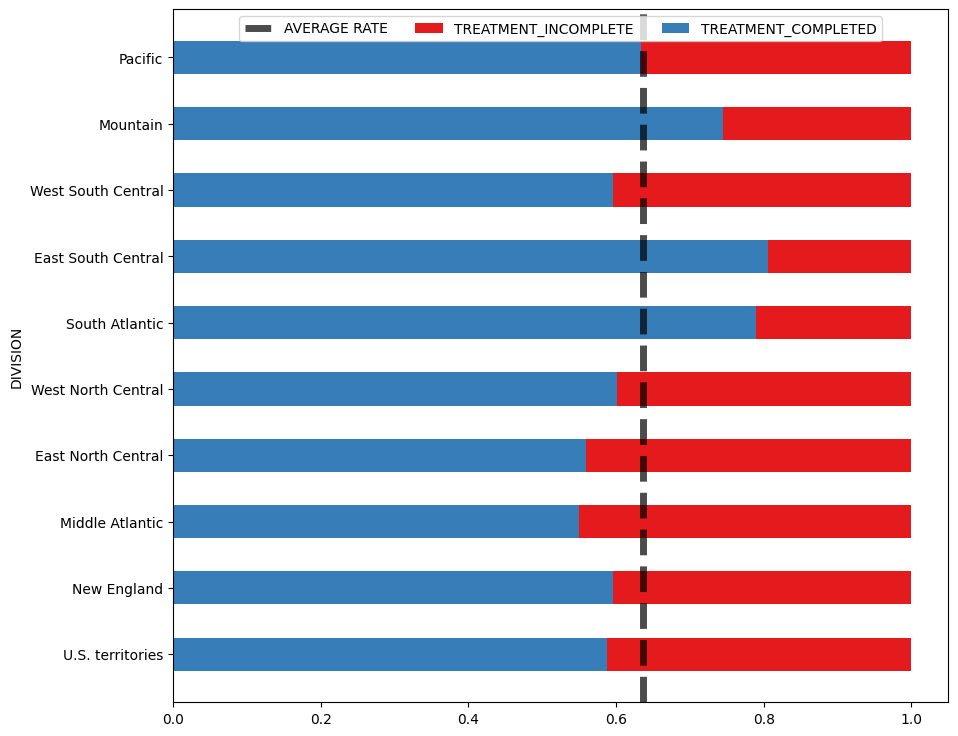


Figure 2‑6 Division on Success rate

## Successfully treated drug

Non prescribed methadone is the drug that has the highest success rate in the least time. It is shown in. Figure 2‑7 and Figure 2‑8.

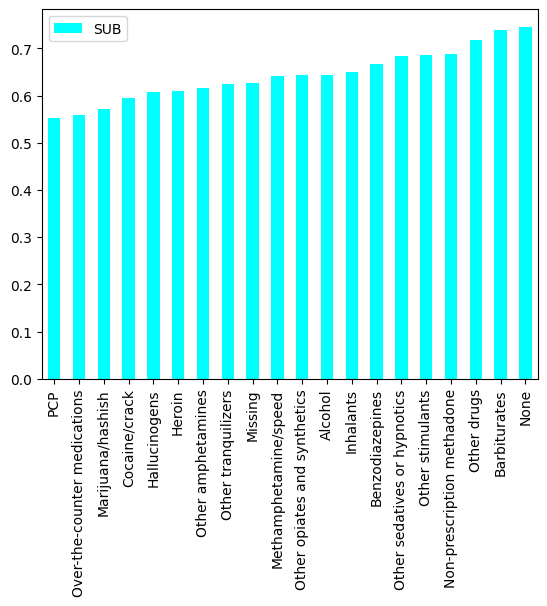


Figure 2‑7 Success probability for diff drugs

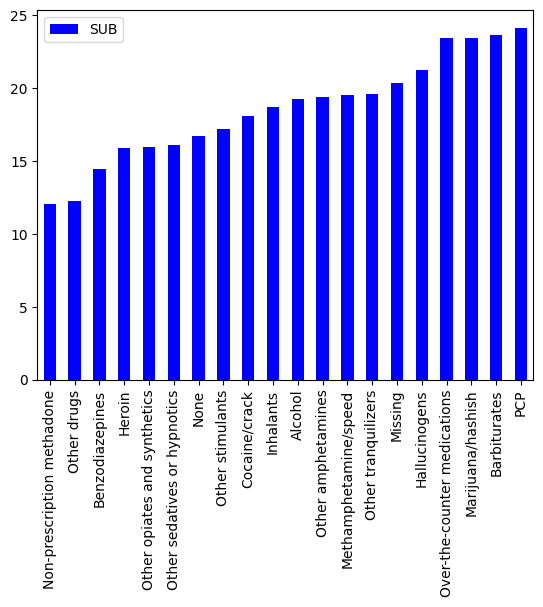


Figure 2‑8 Average time for the success of diff drugs

## State wise variation.

Some of the state governments are doing a great job by establishing more rehabilitation centers. So, we checked no of people (based on population) per rehab center. As shown in Figure 2‑9 . States like New Mexico is doing good as the government had taken initiative to establish more centers for the population. But states like Georgia is not working on this subject.

A picture containing text, plot, screenshot, line

Description automatically generated

Figure 2‑9 Ratio of population to rehab centers.

# MODEL

We have converted the problem to be a binary classification problem. We tried two different models starting with basic logistic regression. Algorithm. It gave us a descent accuracy of 70%. Considering the data is not biased and our wrong prediction doesn’t harm anyone, so we decided to proceed with accuracy to measure our metrics parameter.

# CONCLUSION AND RECOMENDATION

## Recommendation Regarding Assumption

This analysis assumed that the transferred people are successfully treated, which is not evident. So, it should have been a multiclass classification problem instead of binary class. After studying transfer behavior like why they are getting transferred we can classify them further if it was successful or unsuccessful treatment.

## Model Improvement