Project Proposal

Title: Predicting the likelihood of relapse for select addictive substances following rehabilitation

Team members: Yeninda Tjoa, Erica King, Cathy Wang

Subject Area: Public Health

Team name: Sober Signals

Proposed Research

As substance abuse and drug-related deaths grow more prevalent in the United States, rehabilitation is at the forefront of providing treatment for those suffering from addiction. We aim to explore the efficacy of rehabilitation as we hypothesize that the effectiveness of treatment directly impacts the likelihood of relapse. This project aims to analyze data on substance users who have undergone rehabilitation in order to explore the relationship between the users' past involvement with rehabilitation programs and their substance usage and predict the likelihood of relapse following treatment.

Data Explanation

We will be using 2022 and 2023 data from the National Survey on Drug Use and Health which measures substance use, mental illness and treatment in ages 12 or older. One additional analysis we plan on doing is finding a subset of respondents who have participated in this survey in both 2022 and 2023 to determine if there is any progress in their recovery.

Technical Plan

We are considering two approaches: either solely a logistic regression model or beginning with clustering prior to the logistic regression model. The approach we use will depend on whether we can successfully define a response variable from our dataset. If we have a response variable, we will use a logistic regression model to train and test our data. We plan to determine the features we will be using after we clean our data and compute the coefficients and the associated p-values of the features to conclude which ones are significant. Otherwise, we will use clustering to identify patterns in the data to determine the factors for our response variable. Some clustering methods we are considering are K-Means, Gaussian Mixture Models, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Once we define the response variable, we will use a logistic regression model to estimate the likelihood. We plan to use cross-validation to evaluate the performance of the model where we repeat the training and testing process multiple times using a different part of the dataset each time for testing.

Potential Impact

The results of this project will help rehabilitation programs recognize high risk individuals and tailor their treatment plans to effectively help all types of substance users. We hope the outcome can suggest improvement in long-term rehabilitation strategies and increased monitoring of substance abusers.

References

National Survey on Drug Use and health. SAMHSA.gov. (n.d.). https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health