**DS Project**

**Names: Shady Yasin & Mohammad Nujedat & Elaad David :)**

**Topic: Predicting Substance Abuse Treatment Success: Drug Type and Demographics Analysis.**

**Links:**

#### - Data folder: https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/TEDS-A-2015-2019/TEDS-A-2015-2019-datasets/TEDS-A-2015-2019-DS0001/TEDS-A-2015-2019-DS0001-bundles-with-study-info/TEDS-A-2015-2019-DS0001-bndl-data-r.zip

#### - Dataset Documentation:

#### https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/TEDS-A-2015-2019/TEDS-A-2015-2019-datasets/TEDS-A-2015-2019-DS0001/TEDS-A-2015-2019-DS0001-info/TEDS-A-2015-2019-DS0001-info-codebook.pdf

#### Code Link(GitHub):

#### 

**1- Introduction:**

The general problem area this analysis addresses is improving substance abuse treatment outcomes, specifically predicting treatment completion based on drug type. Successfully completing treatment is crucial for individual recovery and public health. Predicting treatment completion is important because it helps tailor interventions and optimize resources. Previous research indicates drug type significantly impacts treatment outcomes. This task is challenging due to the complex relationships influenced by biological, psychological, and social factors, and variability among different demographic groups.

**Data Science Questions:**

**Exploratory**: What are the associations between drug type and treatment completion rates across demographics? We will use logistic regressions to identify correlations.

**Predictive:** Can we predict treatment completion likelihood based on drug type and demographic factors? We will use feature engineering and machine learning models for prediction.

**Hypothesis:** We hypothesize that certain drugs will exhibit higher completion rates. For instance, drugs with milder withdrawal symptoms may correlate with higher success rates. We anticipate significant correlations among drug type, demographics, and treatment completion. Our approach involves cleaning the TEDSA\_PUF\_2015\_2019 dataset because these years offer a comprehensive and recent snapshot of substance abuse treatment trends and outcomes, ensuring our analysis reflects current practices and challenges in the field. We are applying logistic regression and machine learning models to analyze patterns and provide actionable insights for improving treatment strategies.

**2- data overview:**

The TEDS-A dataset includes almost 10 million records, each representing an admission to substance abuse treatment. The dataset encompasses 64 features, which can be broadly categorized into two main groups:

-**Admission Demographics:** Age, sex, race/ethnicity, employment status, and marital status, Location details such as state (stfips) and county (cbsa2010), and Other demographic features like education level (educ) and veteran status (vet).

**-Substance Use Characteristics:** rimary, secondary, and tertiary substances used (sub1, sub2, sub3), Age at first use (frstuse1, frstuse2, frstuse3), route of administration (route1, route2, route3), and frequency of use (freq1, freq2, freq3), Criminal justice system involvement (detcrim, arrests), and Treatment-related factors like the number of prior admissions (noprior) and referral source (psource).

Each record also includes unique identifiers and various health and social indicators, providing a comprehensive view of the treatment admission profile.

3- **Methods & results:**

**-Data Preparation and Preprocessing:** The TEDSA\_PUF\_2015\_2019 dataset underwent thorough preprocessing to handle missing values and standardize categorical variables such as gender, race, ethnicity, employment status, and health insurance coverage. Imputation and categorization ensured data integrity across demographic and substance use variables.

**-Exploratory Data Analysis (EDA):** Initial exploratory visualizations (Figures 1 to 4) provided insights into the distribution of key variables among participants. These visualizations revealed demographic trends such as age, gender, ethnicity, and employment status, health insurance, as well as substance use patterns across primary (Sub1), secondary (Sub2), and tertiary (Sub3) substances. Notably, Figure 5 highlights the completion rates across different substance types, setting the stage for predictive modeling.

-**Figure 1:** Demographic Distribution:

-Gender: Balanced representation.

-Age: Concentration in the 19-25 range.

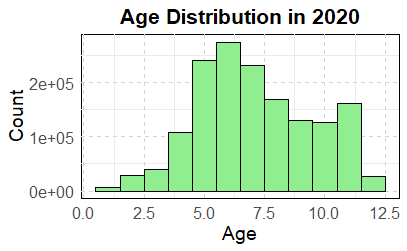
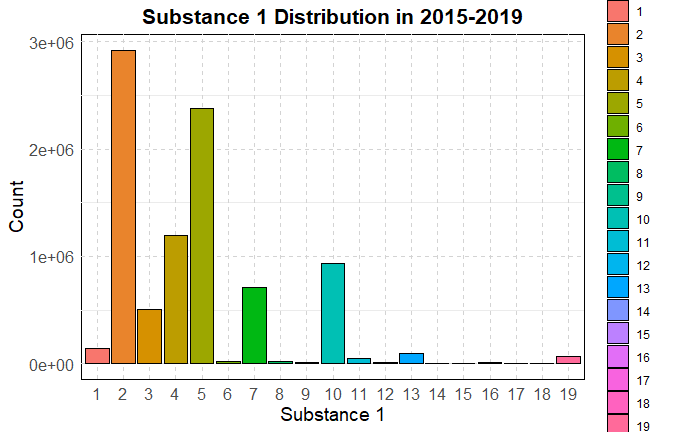
-Ethnicity: Diverse representation.

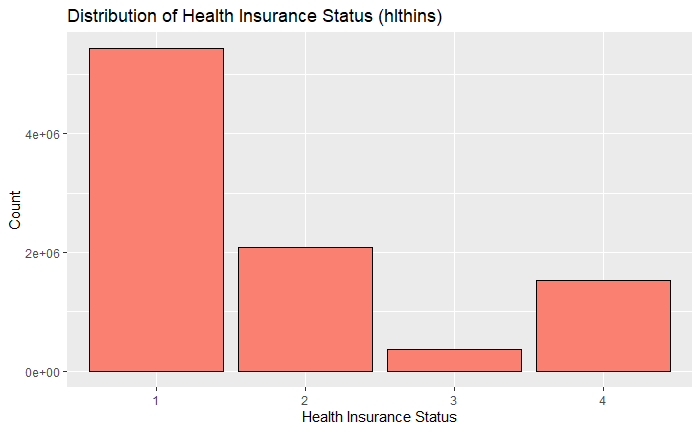
-Employment Status: Varied distribution.

-**Figure 2:** Substance Use Distribution

-Sub1, Sub2, Sub3: Frequency distributions revealing prevalent substances:

Here is some visualization for the most effective based on the p-value:



****

**Feature Engineering:**

To enhance model performance, additional features were engineered:

-Age Groups: Categorized participants into logical age brackets.

-Completion Status: Binary variable indicating treatment completion.

**Model Selection: Logistic Regression**: A logistic regression model was chosen due to its interpretability and suitability for binary classification tasks, aligning with our goal to predict treatment completion based on demographic and substance use factors.

**Model Training and Performance:** The logistic regression model was trained using predictors including Sub1, age, gender, ethnicity, race, employment status, education level, health insurance coverage, psychiatric problems, and frequency of substance use (freq1, freq2, freq3). Model evaluation on the test set yielded the following metrics:s

-Accuracy: 0.63

-Sensitivity: .511

-Specificity: 0.727

-AUC: 0.686

Accuracy: 0.63

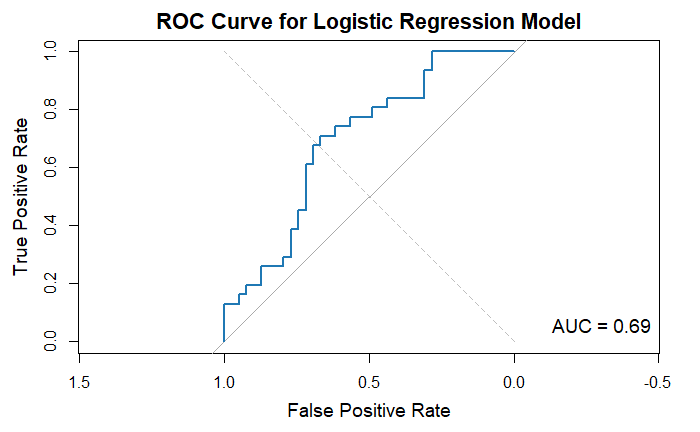
Sensitivity (True Positive Rate): 0.511

Specificity (True Negative Rate): 0.727

Setting levels: control = 0, case = 1

Setting direction: controls < cases

AUC: 0.686

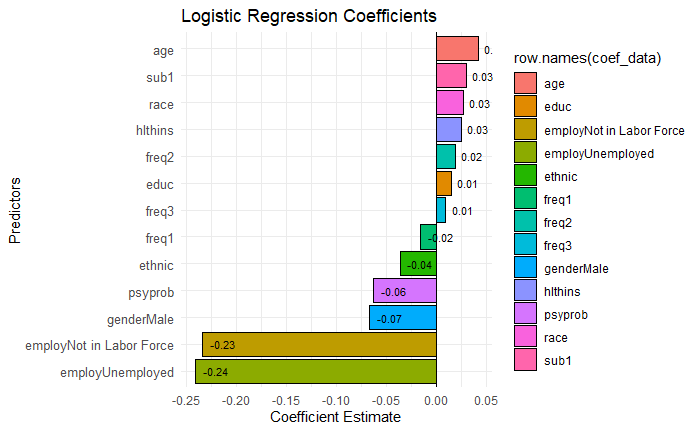


**Interpretation and Findings:**

The logistic regression coefficients (Figure 6) indicated significant predictors of treatment completion:

-Positive Coefficients: Associated with factors positively influencing completion.

-Negative Coefficients: Associated with factors negatively influencing completion.



**Model Validation:** Cross-validation and ROC curve analysis (Figure 7) validated the model's robustness and predictive capability, demonstrating consistent performance across different folds and confirming the model's fit to the data.

**Discussion:**

Implications and Limitations: The results underscore the predictive power of demographic factors and substance use characteristics in determining treatment outcomes. Insights gained from EDA and model findings highlight opportunities for targeted interventions based on demographic profiles and substance preferences to enhance treatment efficacy and completion rates. However, limitations include potential biases from self-reported data and the complexity of real-world treatment dynamics not fully captured by the model.