Classifying Opioid Abuse Using Machine Learning

Regis University

MSDS 696 | Practicum II

Amy Haven Maddox

Project Motivation

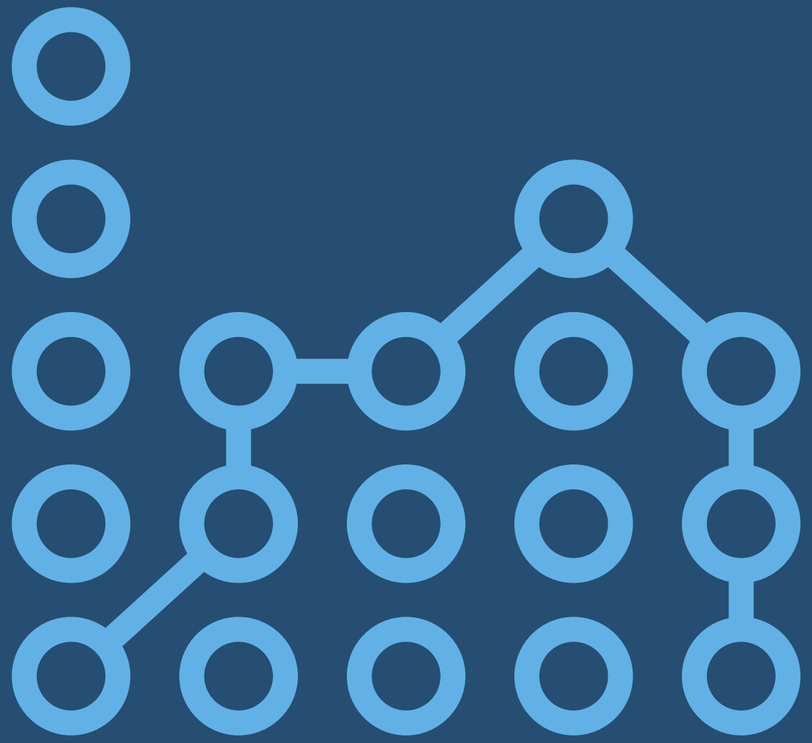
- Since 1999, over **750,000 drug overdose deaths** have occurred in the United States.
- Two-thirds of those deaths can be attributed to opioid substances.
- Opioid addiction is a national epidemic.

Project Motivation

- The motivation for this project drove my decision to continue working with this data set, even when it was tedious and time consuming.
- The overall intent was to obtain specific demographic information on participants, specifically regarding mental health treatment.

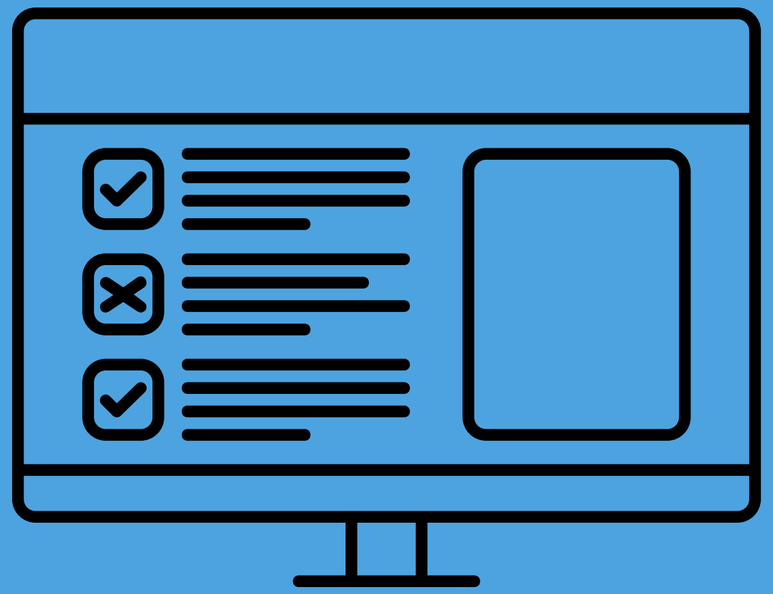
Project Problem

- Can opioid abuse be predicted based on demographic predictors such as age, race, gender, income, and location?
- Does mental health have any impact on whether or not a person will use opioids?



WHAT ARE OPIOIDS?

- An opioid is used to reduce a feeling of pain. The most common forms of opioids are the illegal drug heroin, as well as legal prescription drugs such as oxycodone, morphine, and codeine.
- This project tracks the use of prescription drugs that are misused, as well as heroin.



2018 National Survey of Drug Use and Health (NSDUH)

Retrieval and Sponsorship

- Retrieved from the Substance Abuse and Mental Health Services Administration (SAMHSA) website
- Sponsored by the Center for Behavioral Health Statistics and Quality (CBHSQ)

Target Population

- Civilian, non-institutionalized population of the US over the age of 12 at the time of the survey
- Includes civilians residing on US military bases
- Samples from all states are includes, based on the population

Dataset Information

- Raw dataset includes 56,313 observations (rows) and 2,691 features (columns) of data
- Responses are numerically coded based on the participant's response. The survey uses standard code conventions to determine values for each answer, in general.
    - Reduces the need for data clean-up
    - Handling missing data is straight-forward and simple

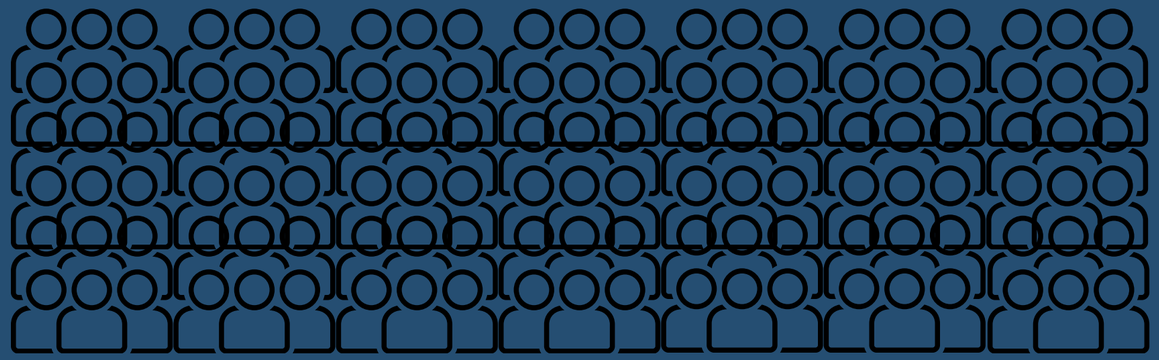
2018 National Survey of Drug Use and Health (NSDUH)

Survey Data Sections

- Self-Administered Substance Use
- Special Drugs
- Risk
- Dependence / Risk
- Special Topics
- Drug Treatment
- Health
- Mental Health (Adult / Youth)
- Youth Experiences
- Consumption of Alcohol
- Demographics

Balancing the Target Variable

- Target variable is "OPINMYR".
- Defined in the Code Book as "Opioid misuse in the past year." Includes both prescription medication misuse and heroin.



53,925

participants report no misuse in the past year

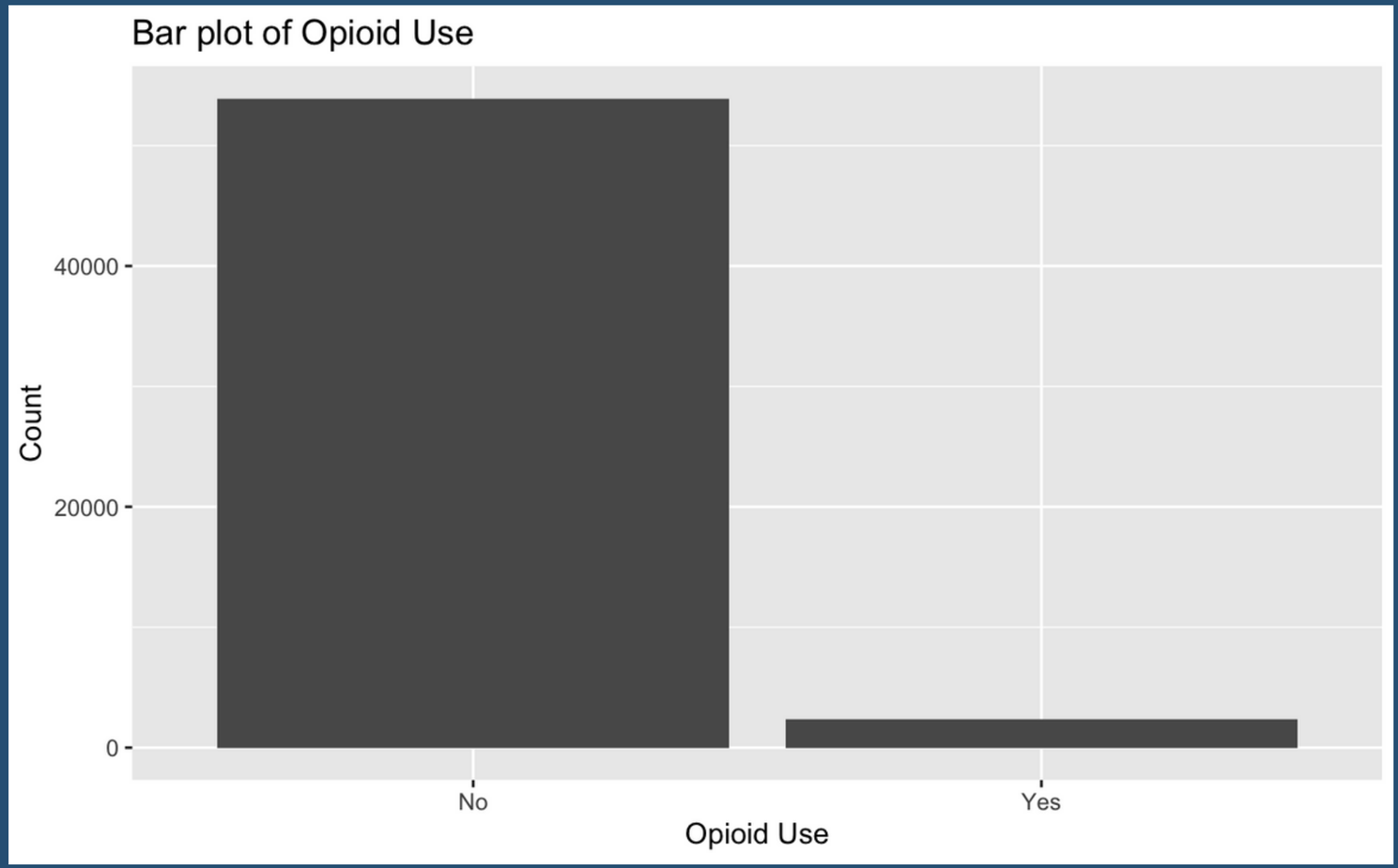
95.7%

2,388

participants report misuse in the past year

0.043%

Balancing the Target Variable



Balancing the Target Variable using Oversampling

Why does the target variable need to be balanced?

- Because the model will use existing data to create a prediction model, there needs to be a balanced amount of participants who have reported opioid use and who have not. Without this, the results from the model will be unreliable.

What is oversampling?

- Observations from the minority class are copied and randomly sampled in order to balance the data.
- For this project, I use the ROSE (Random Over Sampling Examples) package in R, and apply the "ovun.sample" function.
- This creates a new data set with a completely balanced target variable.

Feature Selection

USING THE BORUTA PACKAGE IN R

Why?

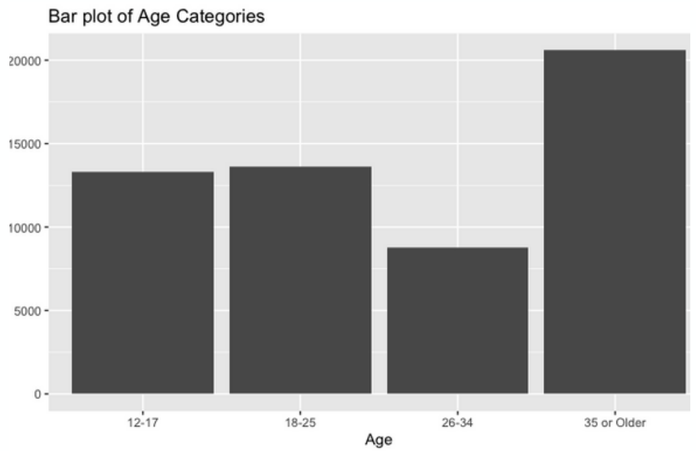
- The NSDUH data set contains 2,691 features.
- Initial data clean-up included removing repetitive data.
    - Even after removing the repetitive data, 866 features remained.

What is the Boruta package?

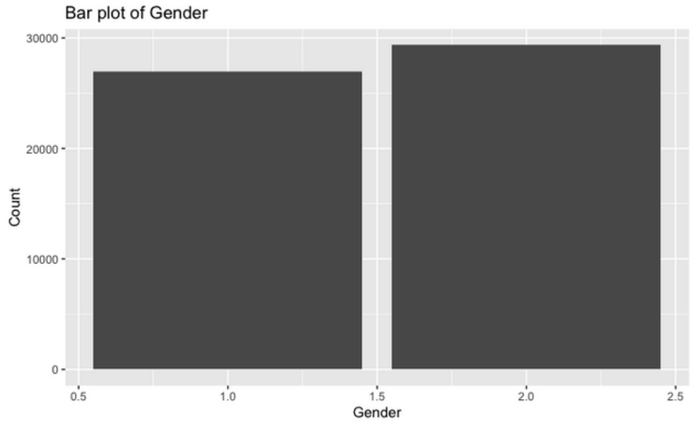
- The Boruta package is a wrapper algorithm and is built around random forest to identify important features.
- I ran the algorithm on the data set twice. Each time, the algorithm ran 99 iterations.
    - There are 66 confirmed attributes, 772 rejected attributes, and 27 tentative attributes. I use these results to create a new dataset to run my model.

Exploratory Data Analysis

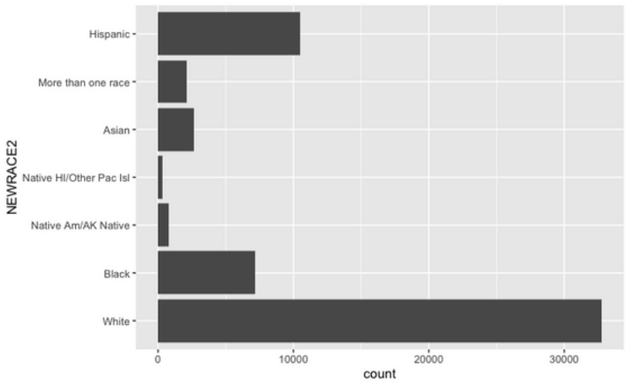
Exploration of Demographic Features



Age Distribution

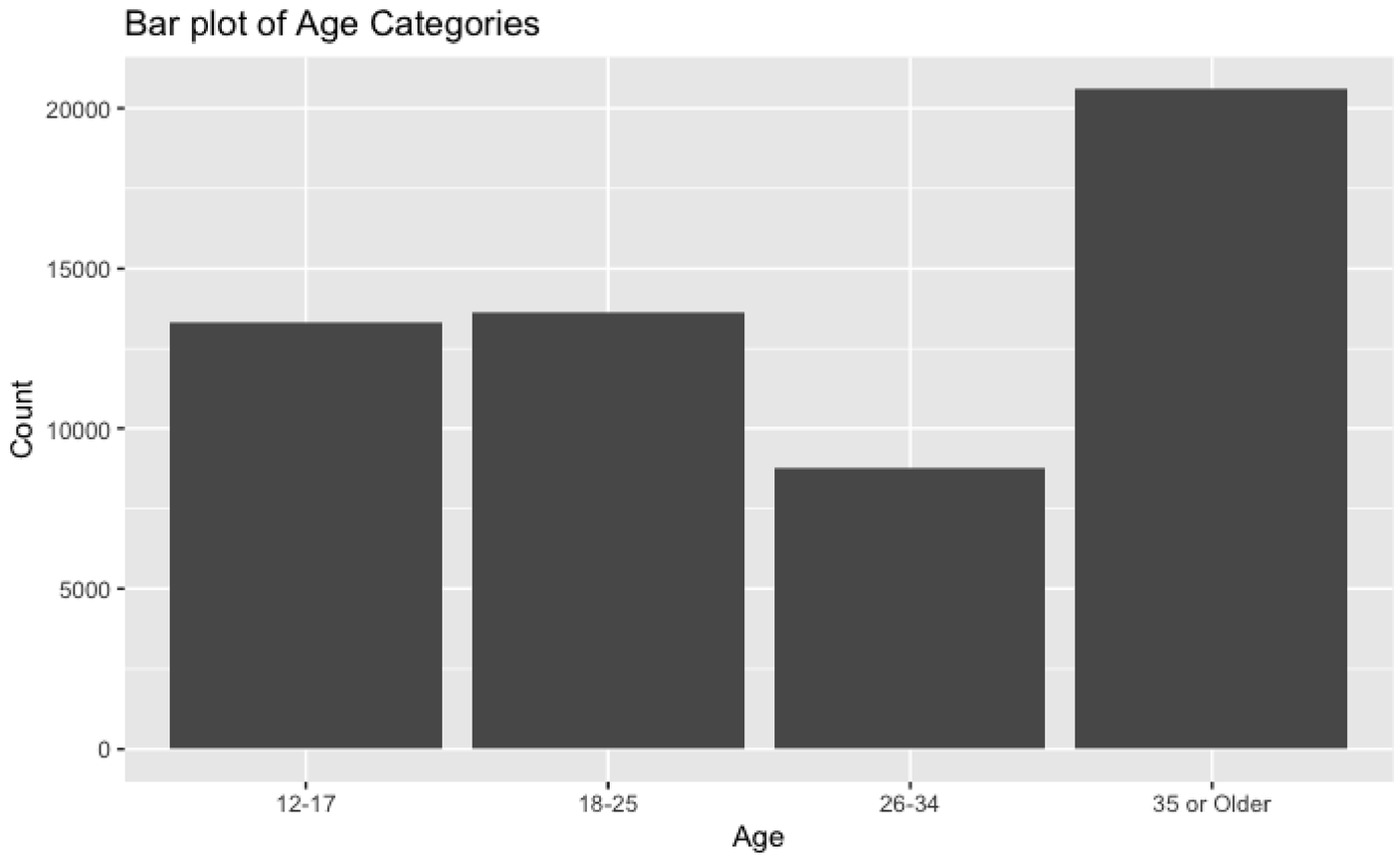


Gender Distribution

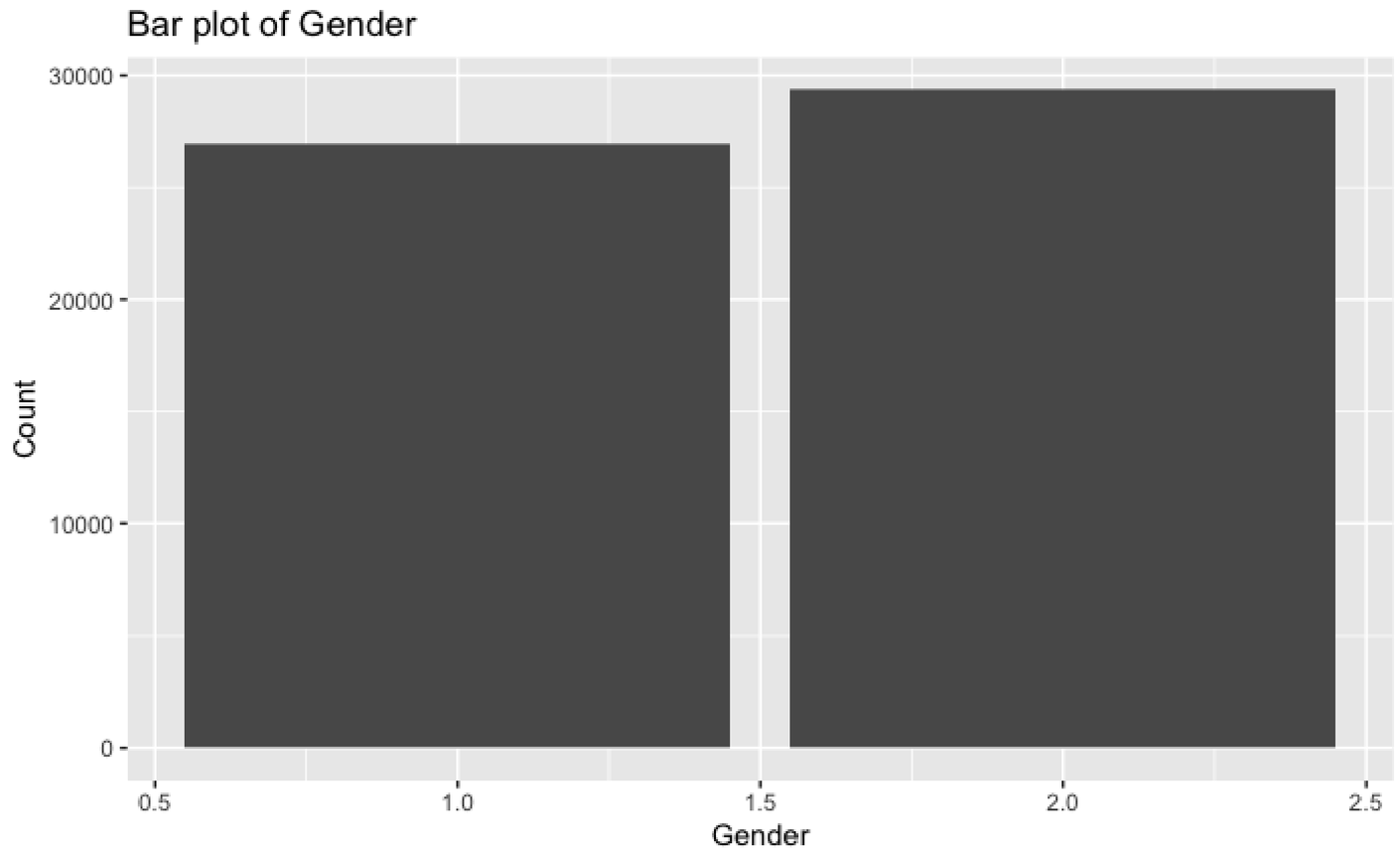


Race Distribution

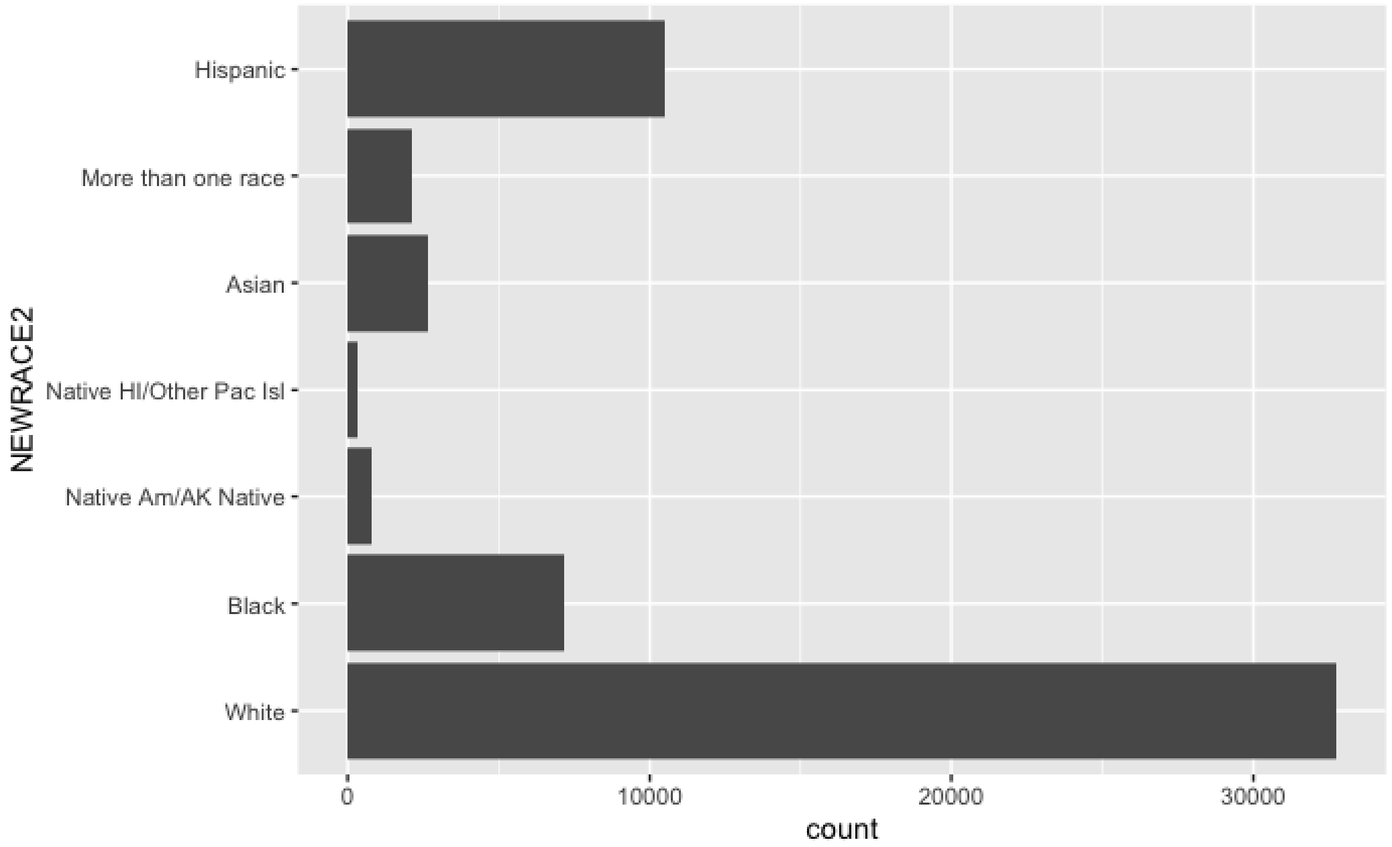
Exploration of Demographic Features



Exploration of Demographic Features



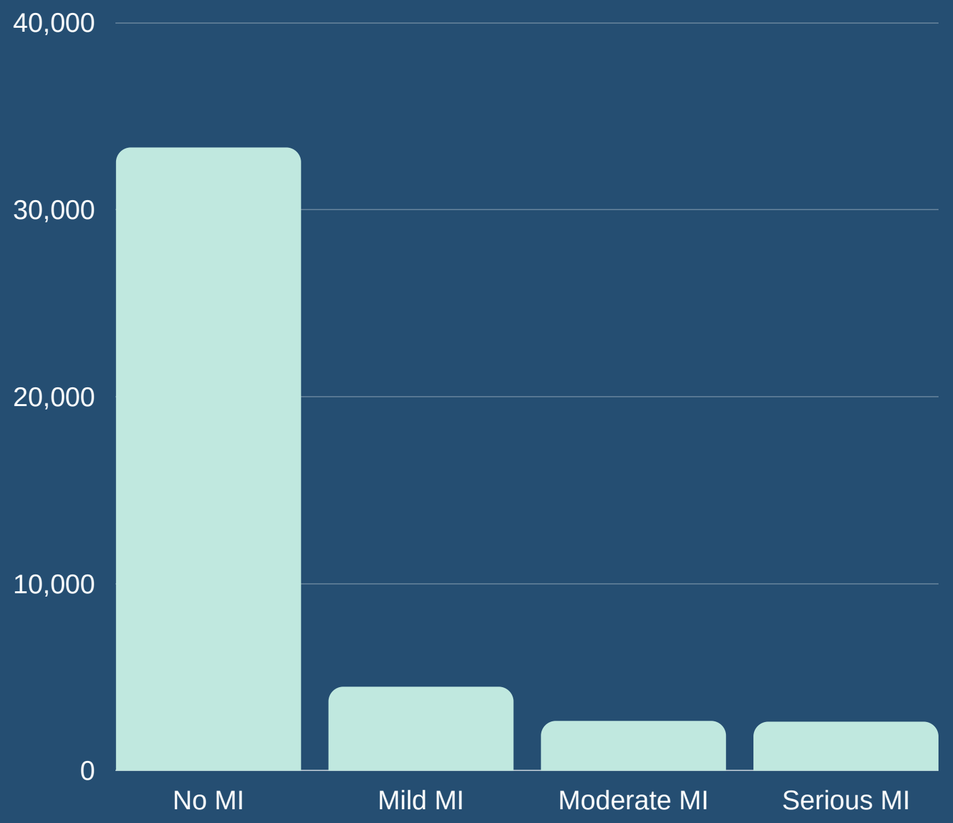
Exploration of Demographic Features



Mental Illness Among Adults

CATEGORICAL MENTAL ILLNESS INDICATOR

- Participants are asked questions throughout the survey regarding their everyday psychological state.
- Responses are coded and inputted into a predictive model based on 2012 data to determine a participants mental illness category.
- Most respondents are categorized in the "No Mental Illness" category.
- "Moderate Mental Illness" and "Serious Mental Illness" volumes are similar, with "Mild Mental Illness" about twice that of the other two categories.



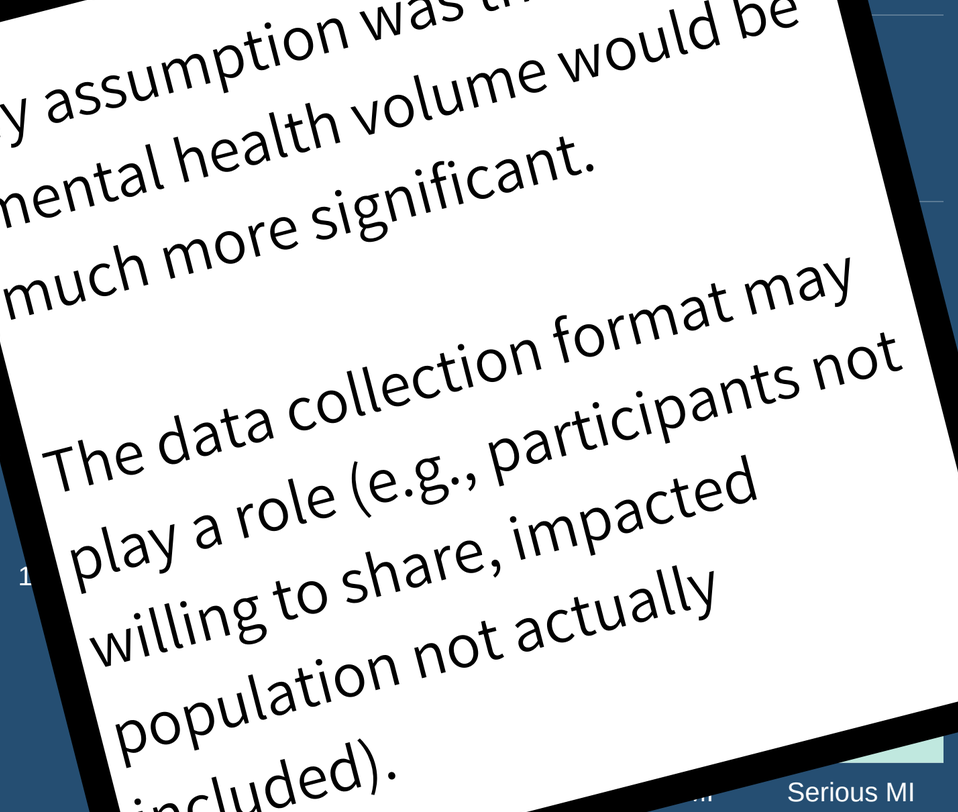
Mental Illness Among Adults

CATEGORICAL MENTAL ILLNESS INDICATOR

- Participants are asked questions throughout the survey regarding their everyday psychological state.
- Responses are coded and inputted into a predictive model based on 2012 data to determine a participants mental illness category.
- Most respondents are categorized in the "No Mental Illness" category.
- "Moderate Mental Illness" and "Serious Mental Illness" volumes are similar, with "Mild Mental Illness" about twice that of the other two categories.

My assumption was that the mental health volume would be much more significant.

The data collection format may play a role (e.g., participants not willing to share, impacted population not actually included).



Building a Machine Learning Model

Implementing the Random Forest

A TALE OF TWO FORESTS

- Originally trained the model on the features deemed important by the Boruta function.
  - 100% accurate
- Ran a second model on a smaller subset of data
  - OOB estimate of error rate of 1.01%

Feature Selection

PART TWO

Why?

- Even with significantly fewer features, the new data set still contains many features, which seems to impact the accuracy of the model.
- What would a model look like with a very small set of features?

How new features are chosen:

- I create a new data set that includes core demographic information on the participants:
  - Age
  - Gender
  - Race
  - Income
  - Work Status
  - Location
- I also include a category for Mental Illness Category and reason for misusing pain medications.

Conclusion and Summary

INCONCLUSIVE

- The data set was not ideal for this project
- I need data that is more clinical, collected directly from facilities

Lessons Learned

CHECK FOR BALANCE.

The data set was imbalanced across multiple features. What methods could have been used to better handle this?

SPEND MORE TIME WITH THE DATA.

I committed to this data too quickly. I should have spent more time vetting the data set before committing.

UNDERSTAND WHERE THE DATA COMES FROM

And how that will impact the result of what I am trying to accomplish. Is survey data the best data for this project?

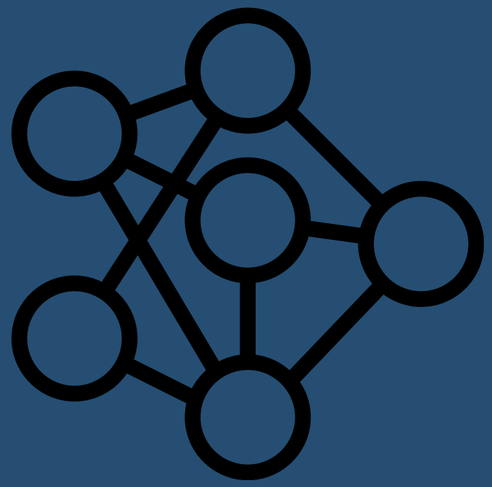
DON'T WASTE TIME

Don't spend too much time on tasks that are not leading anywhere (e.g., feature selection).

Moving Forward

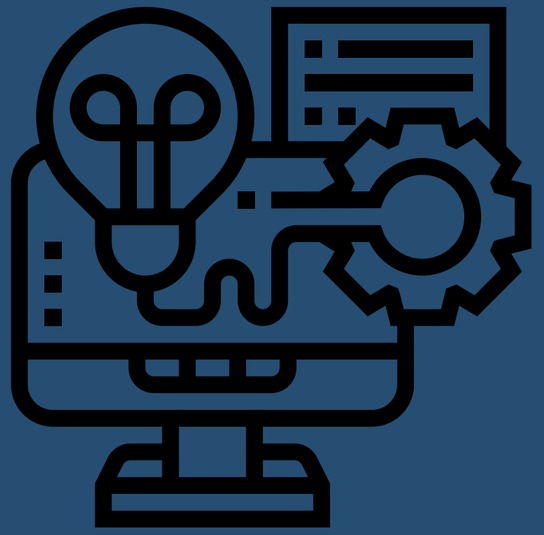
1. Look for more data

Where can I find more suitable data?



2. Re-evaluate the project

Using new data, determine if opioid addiction can be classified using machine learning



3. Expand the project

Is there data on opioid addiction related to the COVID pandemic?

Thank You

Amy Haven Maddox
E-mail Address
amaddox001@regis.edu