Predicting Who Will Suffer from Opioid Abuse

Alicia Abrams
abramsa20@students.ecu

Arup Ghosh
ghosha20@students.ecu

Donna Jarrett
jarrettd21@students.ecu

1 Introduction

The United States of America has a drug problem. From the crack epidemic of the 1980s, to the opioid crisis of today, American citizens have been in a never-ending struggle with substance use and abuse. According to the National Survey on Drug Use and Health (NSDUH), 2 in 5 adult Americans struggled with a substance use disorder.(TBD, 2021) Forty percent of the cause of drug abuse has been linked to genetics. Environmental factors, such as a chaotic home environment, parent's and community attitude toward drugs, peer influences, and poor academic achievement, are found to be just as influential, if not more so. (Hadley, 2021) This leads one to the question of whether future drug use can be predicted by a person's exposure to these environmental factors and more. If healthcare providers and community leaders can anticipate a person's risk for future drug use, then early intervention can take place. Maybe, with proper treatment and community advocacy, we will not see the staggering number of 125,800,000 Americans that struggle with drug use today.

2 Related work

The number of research studies on addiction are limitless. All the articles were found online. Most don't specifically mention their data source however, they're from very reputable sources including The Betty Ford Hazelden Foundation, American Addiction Centers and, National Institutes of Health. Here are a few studies related to our subject:

NIH research article *Opioid Addiction and Chronic Pain* (Emily Petrus and Laura Stephenson Carter, 2018) states, "The United States is facing a double crisis: opioid addiction and unrelieved pain. An estimated two million Americans are addicted to opioids; overdose fatality rates rose more than 20 percent in the past two years. Some 25 million Americans suffer from daily chronic pain and lack effective non-opioid treatments to manage that pain."

Kayla Matthews article *"How Big Data is Changing The Way We Look at Substance Abuse"* (Matthews, 2019) explains how big data can be used to track evidence of prescription drug abuse.

American Addiction Centers research on Genetic and Environmental Factors in Addiction (Centers, 2020) explores the nature versus nurture debate, which is one of the central questions modern science is trying to answer. Why do some people become addicted to alcohol and drugs, but others do not?

The Butler Center for Research at the Betty Ford Hazelden Foundation specializes on addiction research, *Widening the Lens on the Opioid Crisis* focuses on Opioid addiction (Foundation, 2017) They conduct research studies involving patient populations at the Hazelden Betty Ford Foundation. These studies are designed to help identify the mechanisms underlying effective treatment for drug and alcohol problems.

Most individuals who non-medically use Prescription opioids have a history of use. Research evidence supports a high degree of overlap between non-medical prescription opioid drug use and the use of alcohol and other drugs. Therefore, from a prevention perspective, identifying adolescents and young adults who engage in any form of substance use and routing them toward intervention programs will help put the brakes on the opioid crisis.

3 Data

The National Survey on Drug Use and Health provides up-to-date information on drug use, tobacco, alcohol use and other mental health issues in America. The study is conducted every year since 1971 and this year, 70,000 participants were selected. (TBD, 2021) It is the most extensive and most reliable study of American drug use. The data

is publicly available in many formats. It can be downloaded from the Substance Abuse and Mental Health Data Archive (SAMHDA) website in SAS, R, ACII, Stata, SPSS and Delimited. SAMHDA also has an analysis tool called Public Access Data Analysis System (PDAS) for data from 2002 - 2019 The PDAS system allows the user to view the study variables, run crosstab analysis and save the data in a CSV file. (SAMHSA, 2021)

The original data contained 2741 columns. We considered the possibility of using IBM Cloud Pak for Data to determine the most relevant columns. Since the survey responses were coded uniquely per column, it would be time-prohibitive to parse through every column and filter out the responses that were yes and no. So, we used SAMHDA's Public Access Data Analysis System (PDAS) Crosstab Creator to narrow down the columns to distinct columns. The columns that represented redundant survey questions were ignored. The resulting quantity of columns was 29 as follows:

1. Criminal History:
    - Ever Arrested And Booked For Breaking The Law
2. Health:
    - Ever Received Alcohol Or Drug Treatment
    - Body Mass Index (BMI)
    - Age
3. Other drug use:
    - Prescription Sedative Abuse In The Past Year - Imp Rev
    - Prescription Tranquilizer Abuse In Past Year - Imp Rev
    - Methamphetamine Abuse In The Past Year - Imputation Revise
    - Inhalant Abuse In The Past Year - Imputation Revised
    - Smoke All Or Part Of Cigarette In Yr Before Last
    - Ever Used Hallucinogens
    - Smoke Cigarettes Regularly Through Day - Imputation Revised
    - Prescription Stimulant Abuse In The Past Year - Imp Rev
4. Home Life:
    - Parents Helped You W/homework Past 12 Months
    - Tell You They're Proud Of Something You'd Done
    - Number Of Faith-Based Act. Participated In Past 12 Months
    - Father Or Mother In Military
    - Family received government assistance
    - Total family income
5. Mental health:
    - Mental Or Emotional Difficulties
    - Needed Mental Health Treatment But Didn't Get It Past 12 Mos
6. Health insurance:
    - Covered By Health Insurance (Not Otherwise Specified)
    - Covered By Medicare
7. School Life:
    - Now Going To School
    - Work At Job Last Week
    - Attended Any Type Of School Past 12 Months

greatest

Details on how we scrubbed the data, and other code used in this project can be found in our repository on Github at (Github-TeamG, 2021)

We created the following five data-sets:

The first data-set contained the rows that indicated **1 (yes)** in the **Opioid Dependence Or Abuse - Past Year** column.

The second data-set contained the rows that indicated **0 (no)** in the **Opioid Dependence Or Abuse - Past Year** column.

The third data-set was not filtered by opioid use but contained the first 20000 entries.

The fourth data-set was created as a join of data-set three (minus the yes responses in the Opioid Dependence Or Abuse - Past Year column) and data-set one.The data was normalized in data-set four. We changed all the data to 1 for yes, 0 for no. All other data (missing responses, etc... were also 0 for no). The **BMI** column was changed where 18.5-25 = 0, ¡ 18.5 = 1 (underweight), ¿ 25 = 2 (overweight/obese). A normal weight range, 18.5-25.5, should have no effect on drug behavior, so it is 0. Being outside of that range was given a 1 or 2. **Age** was changed so that if the person was ¿18 = 1 and ¡18 =2. Data-set four contained 2367 rows.

The fifth data-set is identical to data-set 4 but **Age** was changed to have 0 for youth and 1 for adults. **BMI** was changed to have 0 for healthy weight and 1 for range below 18.5 and above 25.

The data utilized for Test 1, data-set three, using IBM Watson Machine learning model was reduced to 2000 rows and the columns specified above. Data-set three was used, therefore it included people who had or had not consumed opioids in the last year. The data was pared down using R studios read.csv and write.csv function. Then the data was uploaded to IBM Cloud Pak for Data.

Initially the data was pared down in IBM Cloud Pak for Data data refinery section. The entire raw data was imported into the system. Then the data was filtered by the distinct() operation and the chosen column names were inputted. Then the data was filtered further by the filter operation where the column was Opioid Dependence Or Abuse - Past Year and the value of the column was 1 (for those who indicated yes). However, the sampled responses reduced to only one row. It was unclear whether the entirety of the data was being manipulated or only the sample data. After some research, we determined that the filtering was being applied to the entirety of the data. However, it took around 15 minutes to export the data. Therefore, R studio was chosen to filter data as it only took a few seconds. The data utilized for Test 2, data-set 4, was modified with IBM data refinery. The data utilized for Test 3, data-set 5, was also modified with IBM data refinery.

MySQL MySQL is an open-source relational database management system.

IBM Cloud Pak for Data IBM Watson is an Artificial Intelligence that can analyze data and build machine learning models using Python. Its' Data Refinery contains SQL functions that allow the user to scrub the data. Also, it automatically visualizes the data that is input.

R and RStudio R is a programming language for statistical computing and graphics. Rstudio is an integrated development environment that is made for R and contains extensive development tools.

Jupyter Notebook Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning.

Python Python is an interpreted, high-level and general-purpose programming language.It is widely used in Data Science and Machine Learning. There are lots of libraries available in python that implements Machine Learning algorithms.

GitHub GitHub is a code hosting platform for version control and collaboration. It lets you and others work together on projects from anywhere.

4 Method

Here we will not replicate any similar project or research paper. We used the data set mentioned in the data section. We explored the data and based on that created the data model in relational database. As this data-set has several attributes that are not relevant for our project, we will implement rules during data extraction to cleanse the data. Then we will split the data in training data-set and testing data-set. We will implement below mentioned algorithm to predict the future usage of drugs.

What factors lead to the highest number of opioid addictions? We first filtered the data by people who have indicated that they have depended on or abused opioids in the last year. Then we counted the number of responses in each column to determine which factors are most prevalent in those who use opioids. We will be exploring using IBM Watson to determine the most prevalent factors.

Who will be affected by opioid use? We want to determine whether we can predict whether someone is at risk for future opioid use or dependence. We will be using IBM Watson to implement a machine learning algorithm that will input the data of those already using opioids and compare it to those who are not using it. The more factors they share with those who use it, especially the factors that are most prevalent, the more likely that person will abuse opioids at some point in their life. Currently, we will be using the AutoAI feature of IBM Cloud Pak for data to implement a machine learning model to make predictions about who will likely use opioids. Depending on the results of the tests, the algorithm and Machine learning model will be finalized as the project progresses.

What baseline algorithms will you use? During Data Modelling we will identify the different variables that influence the future usage of drugs. We will separate the variables/factors whether they are discrete or continuous in nature. Based on nature of variable we will implement Baysian network to predict the future usage of drug. In addition, we will explore the use of IBM Watson AutoAI feature to make the prediction. Tests were run to determine whether the model and algorithm would be appropriate. Test 1 and 2 utilized the Random Forest Classifier algorithm.

Test 1: Can we create a Machine learning model based off of data-set three? Test 1

was run in IBM Cloud Pak for Data to determine how to use the AutoAI function and to test if a machine learning model that was generated by the pared down data of 29 columns and 2000 rows would be sufficient to make accurate predictions. We chose the AutoAI binary classification for the AutoAI experiment where we set the prediction column to Opioid Dependence Or Abuse - Past Year. We tried the binary classification instead of the multiclass classification or Regression because Opioid Dependence Or Abuse - Past Year (UDPYOPI) only has two options. Binary classification predicts the likelihood of person falling into 1 or 0 for UDPYOPI. The optimized metric was set to accuracy and the positive class was set to 1. The top performer was pipeline 3 with the algorithm, Random Forest Classifier, a 0.995 holdout accuracy (optimized) and a run time of 00:00:19. The enhancements were 1st hyperparameter optimization and feature engineering. Figure 1 is the Model Evaluation that provides measures of overall predictive accuracy of the model. Figure 2 is the Confusion matrix that shows numbers and proportions of correct and incorrect classifications for the model. Figure 3 shows the progress map that illustrates the steps for creating model pipelines. The chosen pipeline 3 is labeled as P3 and the node is marked with a star.

Test 2: Will normalizing the data create an accurate Machine Learning Model? Test 2 was run in IBM Cloud Pak for Data to determine if an accurate machine learning model could be generated from the data if the survey responses were normalized. The data was modified in data refinery to create data-set 4. A job was created to transform the data from refinery flow to asset. As in test 1, we chose the AutoAI binary classification for the AutoAI experiment where we set the prediction column to Opioid Dependence Or Abuse - Past Year.

Test 3: Will normalizing the data to contain only 1 or 0 responses create an accurate Machine Learning Model? Test 3 was run in IBM Cloud Pak for Data to determine if an accurate machine learning model could be generated from the data if the survey responses contained only 1 (yes) or 0(no). The data was modified in data refinery to create data-set 5. A job was created to transform the data from refinery flow to asset. As in test 1 and 2, we chose the AutoAI binary classification for the AutoAI experiment where we set the prediction column to Opioid Dependence Or Abuse - Past Year.

Test 1: Running the Model against data-set 3 We promoted Pipeline 3 as a model into the deployment space. Pipleline 3 is indicated by a star in figure 1. data-set 3 was deployed to the same de-

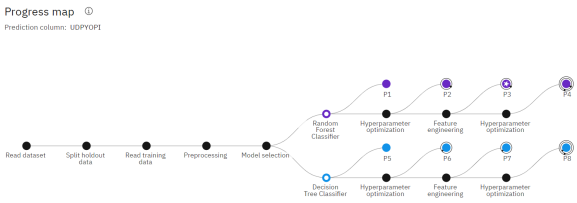

Figure 1: Test 1: Progress Map

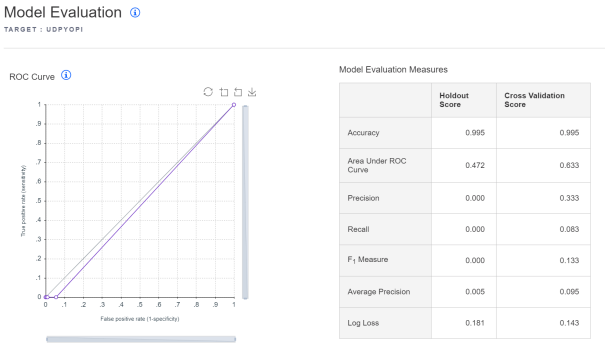

Figure 2: Test 1: Model Evaluation

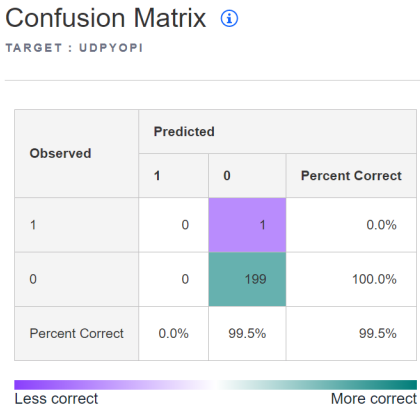

Figure 3: Test 1: Confusion Matrix

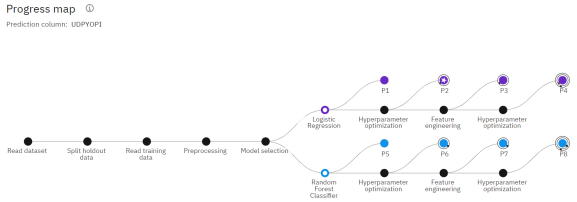

Figure 4: Test 2: Progress Map

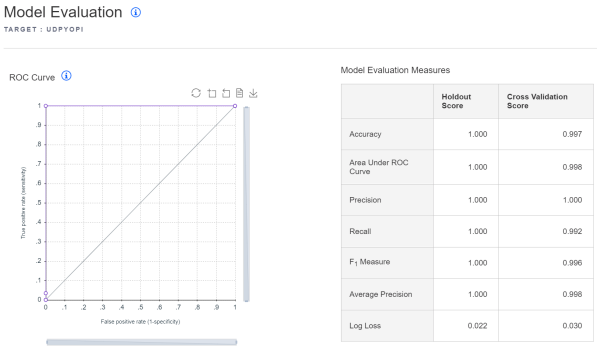

Figure 5: Test 2: Model Evaluation

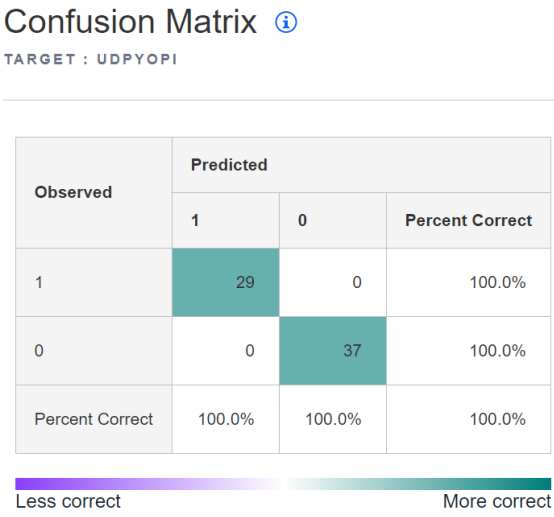

Figure 6: Test 2: Confusion Matrix

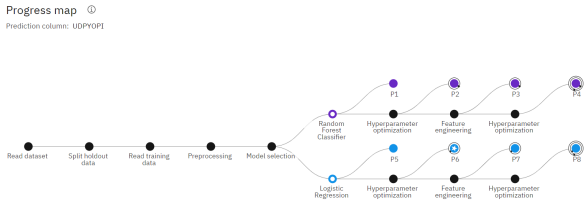

Figure 7: Test 3: Progress Map

Figure 8: Test 3: Model Evaluation

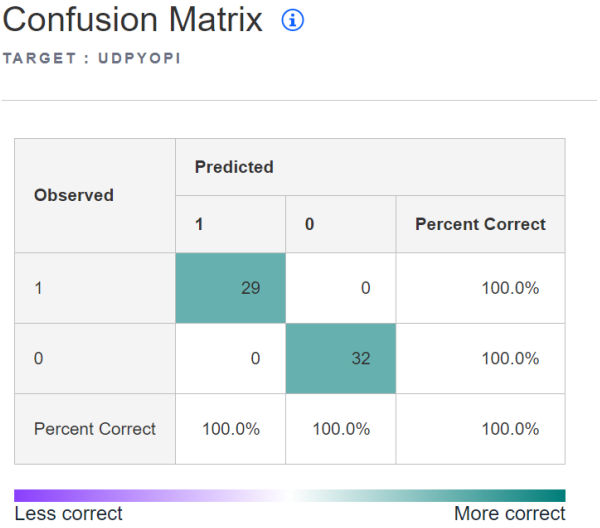

Figure 9: Test 3: Confusion Matrix

ployment space. A job was created and Pipeline3 model was run against the data that was used to create the model. The machine learning service was set as WatsonMachineLearning. The model was run as a batch deployment type. The batch deployment type was chosen because it was to be run on preexisting, uploaded data, not data collected real time from a website. The software specification was hybrid 0.1. The hybrid pipeline software specifications were autoai-kb 3.1-py3.7. The hybrid pipeline hardware specifications were S: 2 vCPU and 8 GB RAM. A job was created and ran. The results were stored in a .csv file that showed the prediction and the probability for each row.

Test 2: Running the Model against data-set 4 Pipeline two was selected to create the model because it was the top performer.It is indicated by a start in figure 4. The holdout accuracy (optimized) was 1.000. The algorithm was logistic Regression. The enhancements were 1st hyperparameter optimization and HPO-1. The build time was 00:00:02. A new deployment space was created. data-set 4 and Pipeline two were deployed to the deployment space. The machine learning service was WatsonMachineLearning. The type was wml-hybrid 0.1 and the software specification was hybrid 0.1. The model was run as a Batch deployment type. Then a job was created to run the model against the data that was used to create the model. The visualization of the results was created with IBM data refinery. We could not use data refinery to visualize test 1 because the data was identical.

Test 3: Running the Model against data-set 5 The training data was 90 percent and the test data was 10 percent. Pipeline 6 was selected

because it was the top performer, it is indicated by a star in figure 7. The holdout accuracy (optimized) was 1.000. The algorithm was logistic Regression. The enhancements were 1st hyperparameter optimization and HPO-1. The build time was 00:00:02. A new deployment space was created to deploy data-set 5 and Pipeline 6. The machine learning service was WatsonMachineLearning. The type was wml-hybrid 0.1 and the software specification was hybrid 0.1. The model was run as a Batch deployment type.A job was created to run the model against the data that was used to create the model. The visualization of the results was created with IBM data refinery.

Progress Map Progress Map for Test-1, Test-2 and Test-3 are depicted in Figure-1, Figure-4 and Figure-7. Progress Map contains all activities and tasks planned or executed during this process. These Progress Maps explains the sequence of the activities (Read Data, Split Data, Read Training Data, Prepossessing, Model selection etc.). Depending on the accuracy of different machine learning model, we need to chose specific for our purpose. As per our test results, Random Forrest Classifier and Logistics Regression are appropriate for our purpose.

Confusion Matrix Confusion Matrix for Test-1, Test-2 and Test-3 are depicted in Figure-3, Figure-6 and Figure-9 respectively. Confusion Matrix is a performance measurement for machine learning classification problem where output can be two or more classes.It is extremely useful for measuring accuracy. As per our Test execution the accuracy is close to 100% for Test-2 and Test-3. Considering the result of confusion matrix, we can conclude that Random Forrest Classifier and Logistics Regression model is accurate for our purpose.

4.1 Milestones & Schedule

1. Discovery Phase (2 weeks)
   - Understanding the Domain (Donna)
   - Understanding the Data (Alicia)
   - Build Use-cases (Arup,Donna,Alicia)
   - Tool selection/Finalization (Arup)
2. Design Phase(2 weeks)
   - High Level Design(Arup)
   - Detail Design(Arup)
3. Implementations(4 weeks)
   - Data Modeling(Arup)
   - Data Extraction(Donna)
   - Data Cleansing(Donna)
   - Creating Tests to determine the best Machine learning model (Alicia)
   - Implementing algorithm on training data set to predict output (Alicia)
4. Analyze the output of the model, do an error analysis (1 weeks)(Arup)
5. Work on final report and presentation (1 weeks)(Donna)

4.2 Results

Test 1: Refactoring the approach to Model Creation The results of test 1 outputted the prediction of 0 and the probability of [1.0, 0.0] for all 2000 rows as shown in figure 10. After further inspection we determined that the sample of 2000 rows only contained three responses of yes. Therefore, the data cannot be pared down by the first 2000 and should be filtered so that the data contains all responses of yes and responses of no to number 2000 total rows.

Test 2: The Data was not Normalized Enough As shown in figure 11, the results do not reflect yes (1) or no (0) responses. This could be the result of BMI and AGE columns containing values other than 1 or 0. The model predicted that 608 would not use opioids and 1759 will.

Test 3: Is this the best Machine learning Model? The data should better reflect more accurate results as the data does not include values other than 0, or 1. The model predicted that 205 people will not use opioids and 2162 will as shown in figure 12.

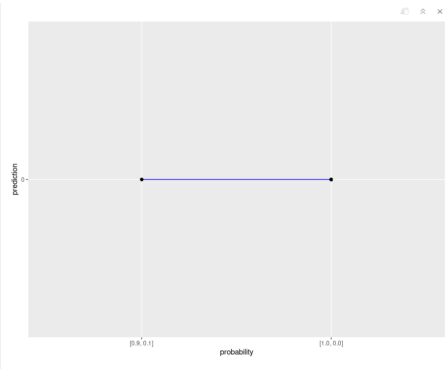

Figure 10: Test 1 Results: Predictions and Probability

Important Features The important features vary by each of the tests as shown in figure 13, and 14. The analysis as to why that is will be discussed in the final project.

References

Centers, E.-S. A. A. (2020). Genetic and environmental factors in addiction.

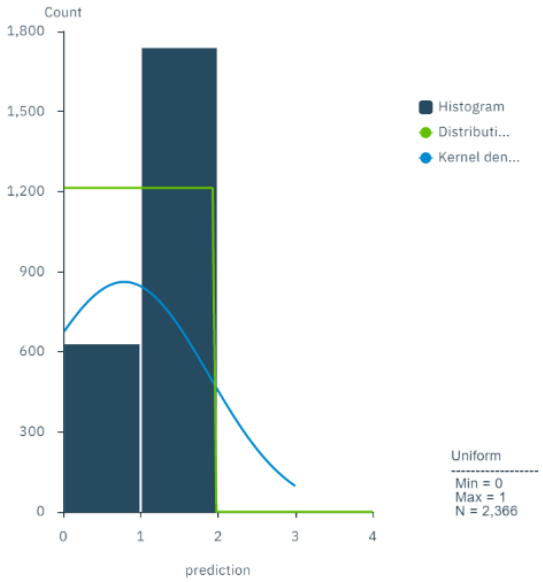

Figure 11: Test 2 Results: Prediction Histogram

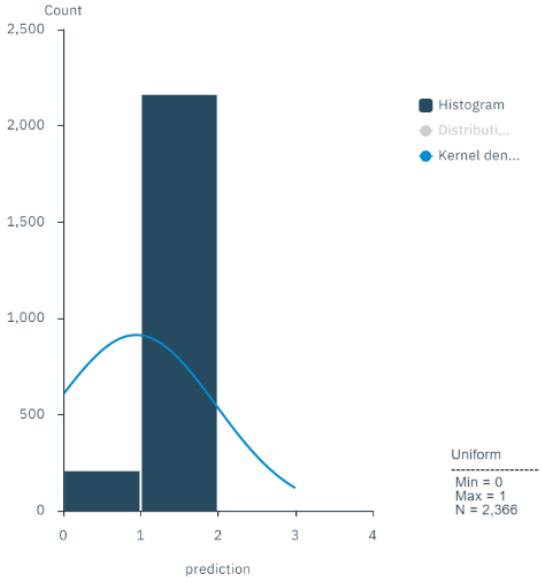

Figure 12: Test 3 Results: Prediction Histogram

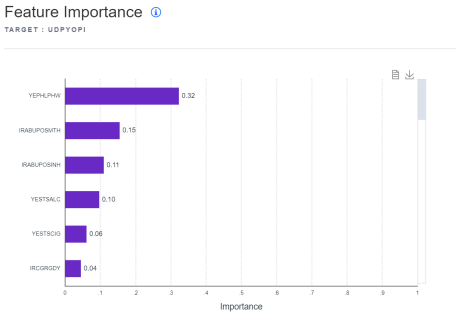

Figure 13: Test 2 Important Features

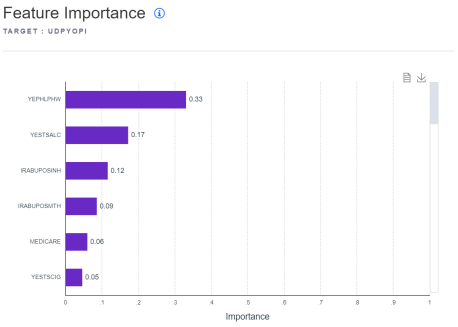

Figure 14: Test 3 Important Features

Emily Petrus, N. and Laura Stephenson Carter, O. (2018). Opioid addiction and chronic pain.

Foundation, H. B. F. (2017). Widening the lens on the opioid crisis.

Github-TeamG (2021). Bigdatafinalproject.

Hadley, S. (2021). Addiction statistics: Drug and substance abuse statistics. retrieved.

Matthews, K. (2019). How big data is changing the way we look at substance abuse.

SAMHSA (2021). Crosstab creator.

TBD (2021). National survey on drug use and health 2019 (nsduh-2019-ds0001).