**Module 8 Portfolio Project**

**Using Predictive Analytics to Analyze a Business Problem**

Joshua Humphreys

Colorado State University Global

Course Code:MIS530

Dr. Osama Morad

October 9, 2022

**Portfolio Project**

In selecting my portfolio project, I decided to try to do work in the realm of addiction. Drug addiction is a topic that is important to me for a variety of both personal and professional reasons. Professionally, addiction has a significant influence on a significant portion of the population that we serve in Emergency Medical Services (EMS). Many or the people that serve have complaints that ultimately can be traced back to a root cause related to a drug addiction. Any work that is done to develop a deeper understanding of drug addiction has the potential to help prevent drug addiction. Drug addiction prevention has obvious humanitarian benefits. The benefits to the healthcare industry as a whole are the “icing on the cake” in the form of increased efficient use of resources.

**Business Problem**

The business problem discussed by my research is whether illicit drug use can be predicted by the use of legal recreational drugs. Legal recreational drugs, for the purposes of this paper, include alcohol, tobacco, and marijuana (despite marijuana being federally illegal for recreational use). The dataset I used comes from the National Survey on Drug Use and Health (NSDUH) and can be found on the Substance Abuse & Mental health Data Archive (SAMHDA). According to SAMHDA (n.d.), “The NSDUH series… is the leading source of statistical information on the use of illicit drugs, alcohol, and tobacco and mental health issues in the United States” (para 1).

The NSDUH surveys the general population 12 years and older. There are many questions related to the use or abuse of many different drugs. In fact, the complete data set contains well over 2000 variables. That is, there are over 2000 questions available in the survey. With that said, the survey does not typically include every question. There are qualifying questions like: have you ever used heroine? If a respondent indicates that they have never used heroine, then they will not have to answer the rest of the questions related to heroine. Additionally, each point in the dataset will be automatically filled with a value that represents the fact that the respondent has never used heroine.

From an analytics perspective, the sheer size of the dataset presents some difficulty. This problem was compounded by a simple problem discovered upon initially loading the dataset in SAS Miner. Most of the questions included in the dataset are categorical. However, when the responses are recorded in the dataset, they are recorded as a numerical code that represents the true answer. For example, A question may be worded as follows: “Please think about how true each statement is of you. After not smoking for a while, you need to smoke in order to feel less restless and irritable” (Center of Behavioral Health Statistics and Quality, 2021, p 229). The answers to that question available to the respondent are “Not at all true”, “Somewhat true”, “Moderately true”, “Very true”, and “Extremely true” (Center of Behavioral Health Statistics and Quality, 2021, p 229). Additionally, as mentioned earlier, there are screener questions that determine, for example, if the respondant has ever smoked a cigarette. If the respondant indicates that they have never smoked a cigarette, then the questionairre won’t ask them any cigarette questions and will automatically fill their answer to those questions in as “Never used cigarettes” (Center of Behavioral Health Statistics and Quality, 2021, p 229). Each one of the possible answers is then given a numerical code which is what gets recorded in the dataset. So, for example, “Not at all true” is recorded as “1”, “Never used cigarettes” is recorded as “91”, and so on (Center of Behavioral Health Statistics and Quality, 2021). The problem is that SAS Enterprise Miner, due to the nature of the numerical codes, recognizes the data as interval type rather than categorical. This wouldn’t be a concern if it wasn’t for the size of the dataset. With over 2000 entries in the dataset, manually changing the type of each imported variable is tedious and prone to error.

To simplify my analysis and data prep, I decided to narrow my question down. I decided that attempting to predict heroine abuse based on alcohol, cigarette, and marijuana abuse would be sufficient for the purposes of this project. In other words, my null hypothesis and alternate hypothesis are as follows:

H­­o: Abuse of alcohol, marijuana, or cigarettes are not predictive of heroine abuse.

Ha: Heroine abuse can be predicted by abuse of alcohol, marijuana, or cigarettes.

**Data Preparation**

Prior to loading the data set in SAS Enterprise Miner, I decided to do a few initial preparation steps in Microsoft Excel. This was primary due to the ease of bulk column actions afforded by the user interface of excel. I isolated the variables representative of cigarette, alcohol, marijuana abuse. I then wrote a formula that created a simple binary heroine abuse indicator named *HERABUSBIN*. *HERABUSBIN* was generated based on all the heroine abuse variables. Essentially, if a respondent indicated a positive answer to any of the heroine abuse variables, a value of 1 was assigned to *HERABUSBIN* for that observation.

Upon completion of the data preparation steps I performed in Excel; I loaded the dataset into SAS Enterprise Miner according to the steps included in the McCarthy text. My first step in analysis was to use the *StatExplore* node. The major takeaways from this initial exploration were that each “superclass” (that is variables related to cigarettes, marijuana, or alcohol) were grouped together in terms of variable worth.

One of my concerns with this dataset is the inherrent correlation between variables. I knew this correlation exists prior to doing any kind of statistical analysis based on my review of the codebook. When I say the variables are inherrently correlated I mean that the questionairre is essentially trying to establish to what extent a respondant abuses or is dependent on a substance. If a respondant has had a cigarette craving in the last 30 days, it is implied that they have had a cigarette craving in the last year. However, the inverse of that situation isn’t true. Furthermore, there is a possibility that the extent to which a person abuses one substance is related to the extent to which they abuse another. That is, someone who abuses alcohol occasionally may be less likely to abuse heroine than a person who abuses alcohol frequently

**Figure 1.**

*Screenshot of dataset in Excel.*

Graphical user interface, application, table, Excel

Description automatically generated

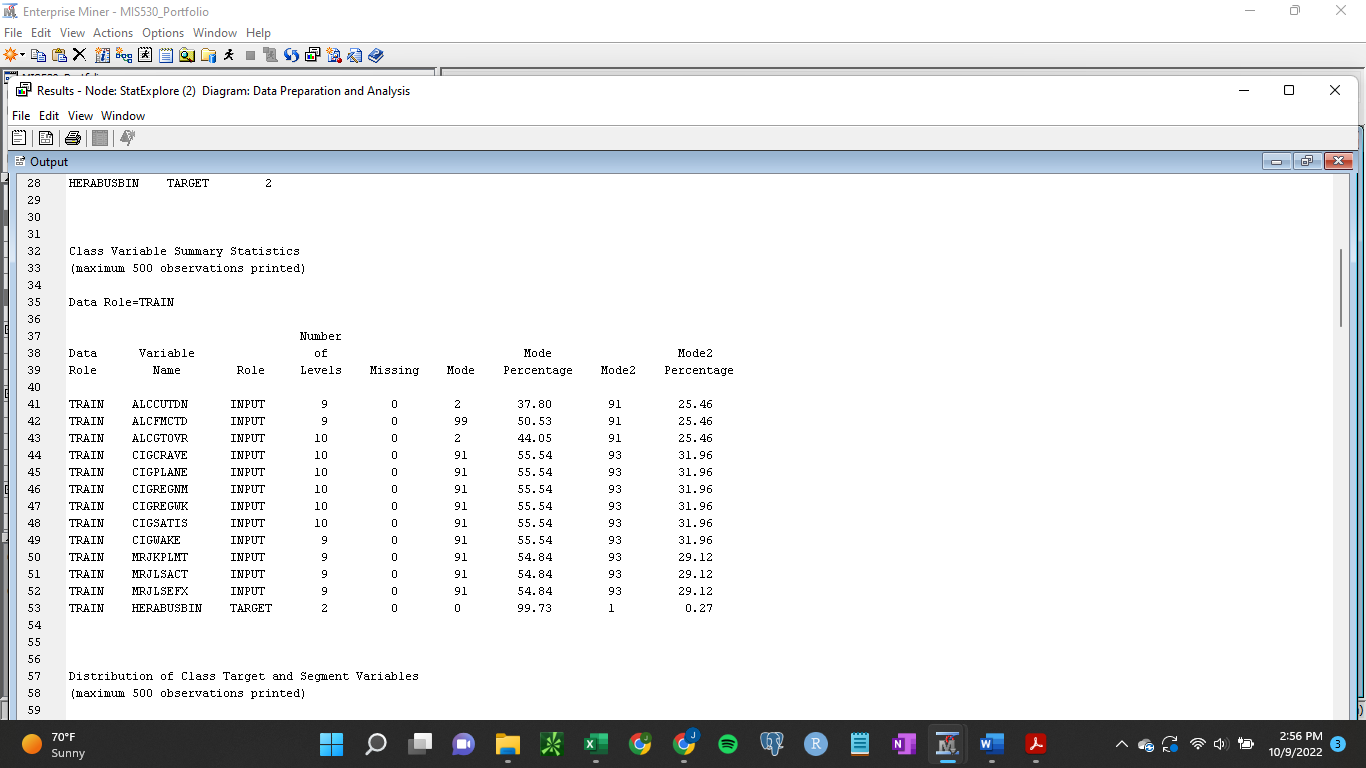
After running the StatExplore node, I decided to run the variable selection node. This was primarily due to the correlation I mentioned before as well as the sheer number of variables. That is, I used the variable selection node to avoid overfitting. The variable selection node was able to narrow the many variables in the dataset down to the few found in Table 1. After running the *Variable Selection* Node, I ran the *StatExplore* node again. The class variable summary statistics from the post *Variable Selection* node data is in Figure 2. The summary statistics from the StatExplore node demonstrates that there are no missing values. Additionally, all the input variables are now class variables. At this point, no other data preparation is necessary

**Table 1.**

*Description of variables used in models***Error! Switch argument not specified.**

**Figure 2.**

*Screenshot of Class Summary Statistics*

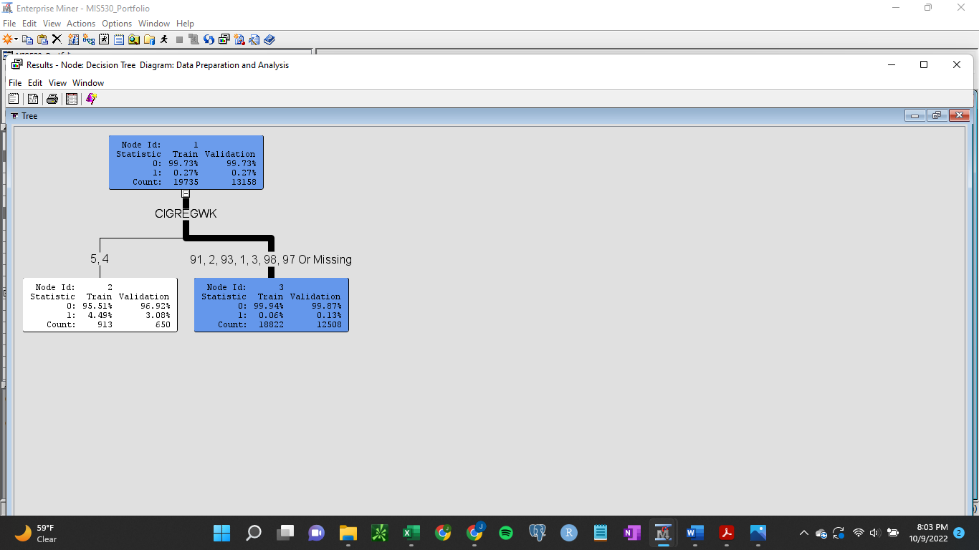
**

**Model Analysis**

I decided that, for this project, the interpretation of the results would be particularly important. That is, the intent of this project is to understand the predictive relationship between legal drug abuse and heroine abuse. Since many of the variables in the dataset provide insight in to varying degrees of abuse, the value of this project could be enhanced if we are able to get an idea as to the level of abuse of legal substances that is (if at all) predictive of heroine abuse. With both decision trees and logistic regression, it’s straightforward to interpret the contribution of each variable to the result. Therefore, I decided to stick with those two models (even though it is also possible to explain a neural net with a decision tree). The results of each of those models can be found in the figures below.

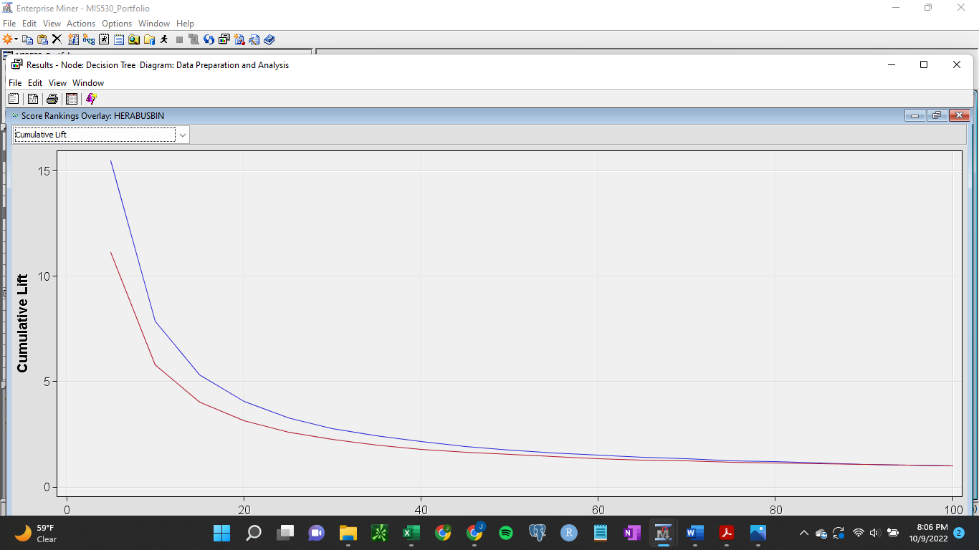
**Figure 3.**

*Screenshot of Decision Tree*

**

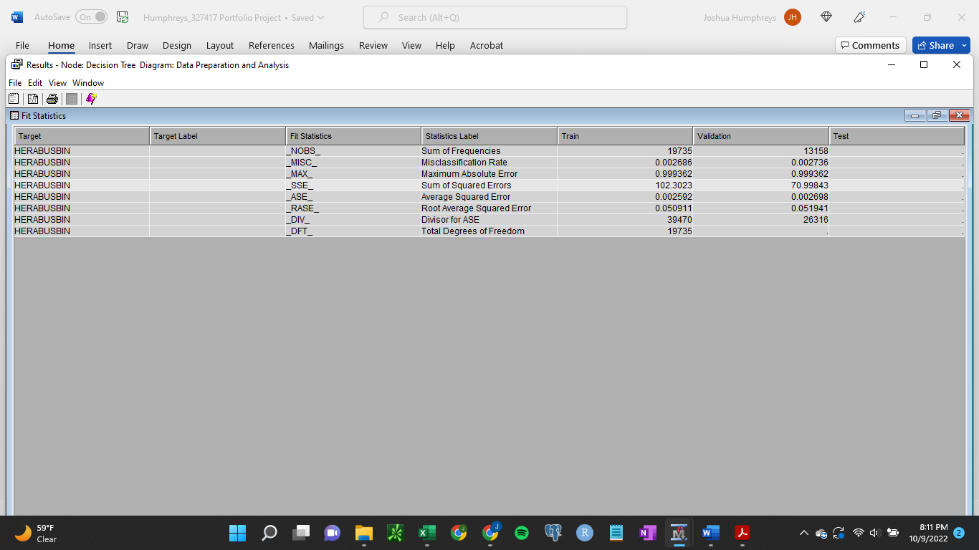
**Figure 4.**

*Screenshot of cumulative lift from decision tree*



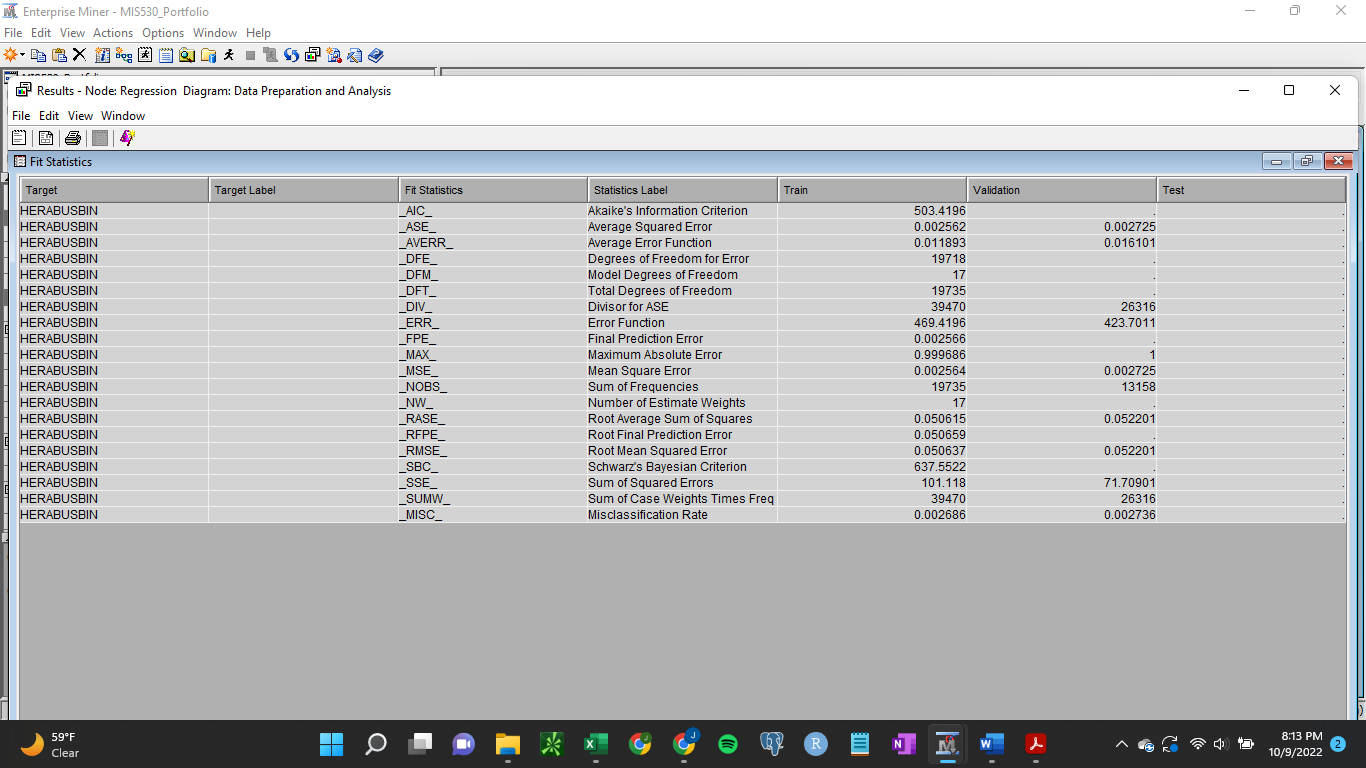
**Figure 5.**

*Screenshot of* *Decision Tree Fit Statistics*

****

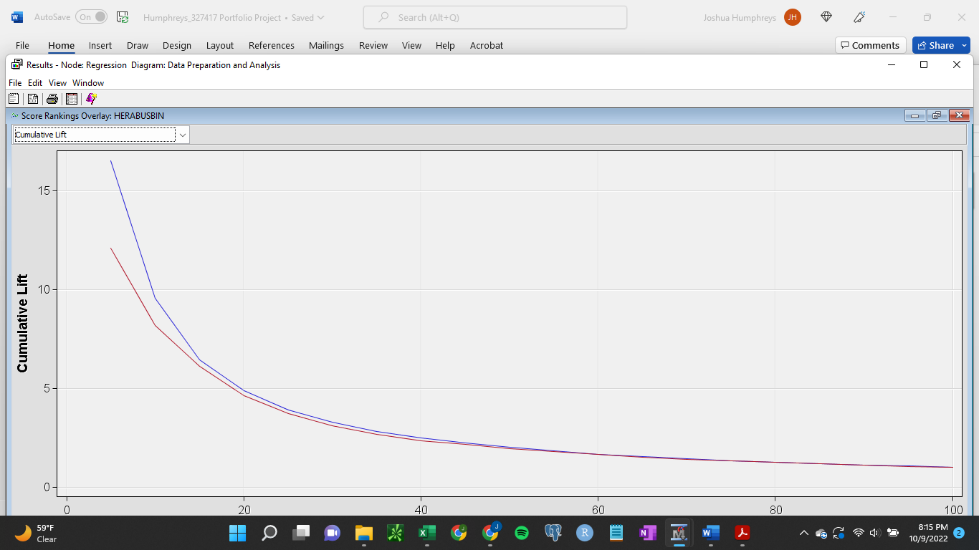
**Figure 6.**

*Screenshot of Regression Fit Statistics*

**

**Figure 7.**

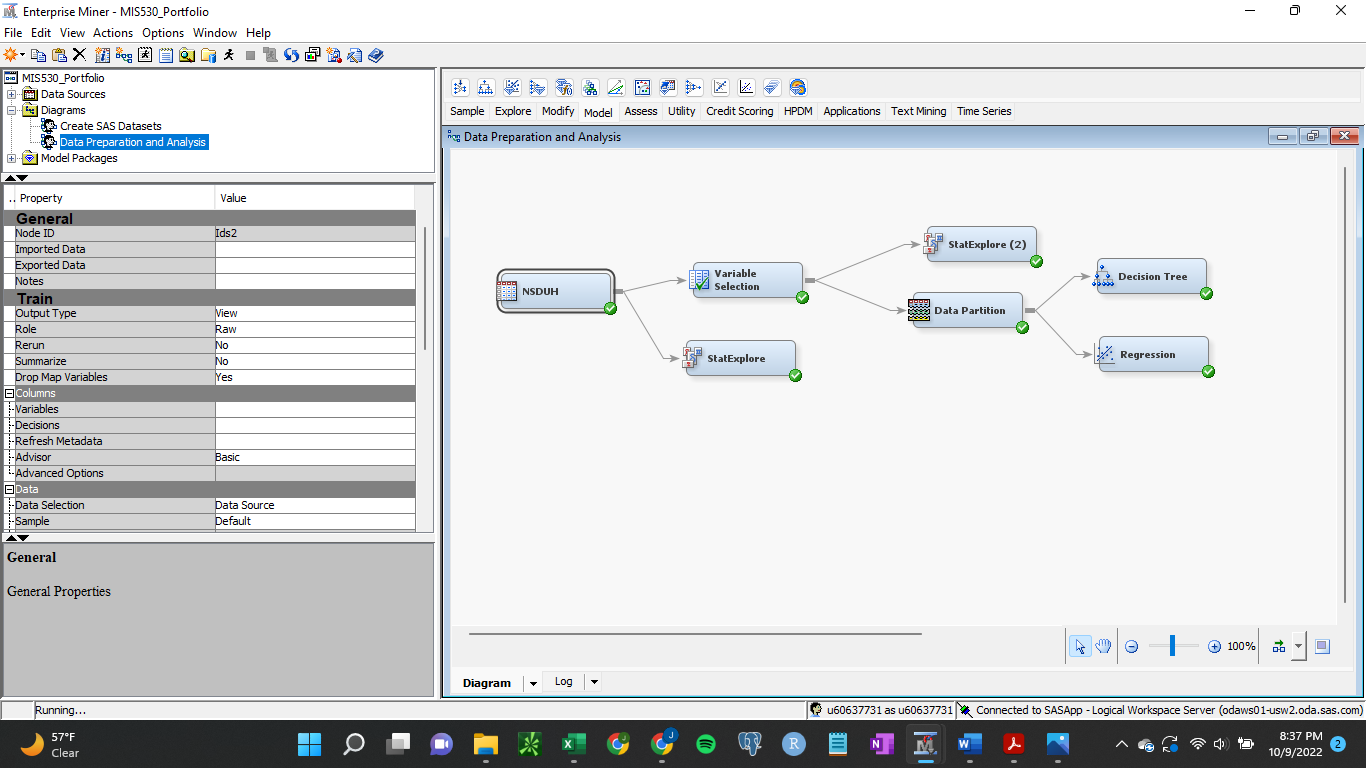
*Screenshot of Regression Cumulative lift*

****

Cumulative lift charts for both models are ok. That is, it appears the model is better than no model for approximately the first 40% of respondents. However, when we examine the decision tree, we see that there are only 3 nodes (root node and two leaf nodes). Examination of the average square error reveals that both models produced an average square error of approximately .0027. When we examine the misclassification rate in each decision node, we see that the decision tree misclassified both the training and validation sets at a rate of approximately .27%. The Logistic regression model misclassification rate was the same. Unfortunately, if we revisit our summary statistics of our dataset, heroine abusers only made up .27% of the respondents. In other words, what is essentially happening is that each model is simply predicting that all respondents are not heroine abusers.

**Figure 8**

*Screenshot of Data Preparation and Analysis Diagram*



**Conclusion**

Based on the results of my analysis, I am unable to reject the null hypothesis. That is, based on my models, I don’t believe that Cigarette, Alcohol, or Marijuana use is predictive of heroine abuse. From a business perspective, this result is not useless. Further analysis could involve isolating only heroine abusers and looking for relationships between variables (possibly using clustering to do so). From a real-world perspective, there are two significant challenges with this type of research that this data set highlights. First, legal drug use is relatively common and heroine abuse is very rare. For these two reasons, it’s unsurprising that my analysis did not lead to the rejection of my null hypothesis

**References**

Center of Behavioral Health Statistics and Quality. (2021). *2020 National Survey on Drug Use and Health Public Use Files Codebook.* Substance Abuse and Mental Health Services Administration. Rockville, MD.

Immortalishika2001 (2021, March 16) *Understanding gain chart and lift chart.* Geeksforgeeks. <https://www.geeksforgeeks.org/understanding-gain-chart-and-lift-chart/>

Glen, S. (n.d.) *Null hypothesis definition and examples, how to state.* StatisticsHowTo.com: Elementary Statistics for the rest of us! <https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/>

Lanhenke, M. (2022, February 7). *Cheat-sheet: Decision Trees Terminology. Finding your way through a random forest of terminology.* Towards Data Science. https://towardsdatascience.com/cheat-sheet-decision-trees-terminology-c37ee6cb2d2e

McCarthy R.V., McCarthy, M.M., & Ceccucci, W.(2022). *Applying Predictive Analytics: Finding Value in Data* (2nd ed.). Springer.

Outside Two Standard Deviations (2018, September 13). *An overview of correlation measures between categorical and continuous variables.* Medium.com. https://medium.com/@outside2SDs/an-overview-of-correlation-measures-between-categorical-and-continuous-variables-4c7f85610365

SAMHDA (n.d.) *National Survey on Drug Use and Health (NSDUH): Population Data.* Substance Abuse & Mental Health Data Archive. <https://www.datafiles.samhsa.gov/dataset/national-survey-drug-use-and-health-2020-nsduh-2020-ds0001>

Sarma, K.S. (2017). *Predictive Modeling with SAS® Enterprise Miner™: Practical Solutions for Business Applications. Third Edition.* Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2017) *SAS® Enterprise Miner™ 14.3: Reference Help.* Cary, NC: SAS Institute Inc.

Wan, A.(2020, April 23). *Making decision trees accurate again: explaining what explainable AI did not.* Berkeley Artificial Intelligence Research. <https://bair.berkeley.edu/blog/2020/04/23/decisions/#:~:text=In%20a%20neural%2Dbacked%20decision,%E2%80%9D%20or%20%E2%80%9Cno%20sausage%E2%80%9D>.