Opioid Addiction | Practicum II

Sam Blumer

August 23, 2019

**Predicting Opioid Addiction Likelihood Using Neural Nets**

The rising use in opioids and opiates has many tributaries, starting as early as the mid-1990’s as the treatment of pain became a priority for doctors while pharmaceutical companies began developing stronger and more concentrated versions of pain medications. According to the Department of US Health and Human Services (2016), this amalgamation fueled the opioid crisis that we know today, leading to millions of individuals misusing prescriptions, ultimately leading to the “opioid overdose deaths of more than 42,000 people in 2016, more than any previous year on record. In fact, an estimated 40% of opioid overdose deaths involved a prescription opioid.”

While the epidemic originally started in smaller midwestern towns, the easy access to the drugs, and the relatively lower prices made the drugs, including heroin, more accessible and used in higher doses. A US HSS study found that the Midwestern region witnessed opioid overdoses increase 70% from July 2016 through September 2017 and Opioid overdoses in large cities increase by 54% in 16 states.

While it is known that opiates are incredibly addicting, it has only been recently studied what characteristics an individual may possesses, whether demographically, biologically, or relative to their drug use, that may lead to addiction. With the recent focus on analyzing data, we are going to attempt to use these indicators to predict the likelihood that an individual may become addicted.

We are going to be using a variety of packages that will allow us to process, explore, analyze, and ultimately offer tools for feature engineering and machine learning. Note that we will also be using a local Spark instance for processing and storing large amounts of our data.

For our local Spark instance, we are going to use Spark 2.1.0 and set-up particular configuration specifications based off the computer we are using for our analysis. This analysis had a basic configuration using 2 cores and 8GB of local RAM to store and process our data via our Spark cluster.

# DATA SOURCES

We have multiple data sources that we are going to inspect and aggregate to make a more complete data set for our final evaluation. In total, we used 11 different data sources. Each data source provided some information related to the demographics, socioeconomic, geographic, cultural, and economic factors for each individual admitted to a rehabilitation facility in the United States in the year 2012 (the most recent data available for public consumption). Included in the contents of this reports includes the [data dictionary and raw data](https://github.com/sam-blumer/OpioidAddictionNN/blob/master/DataDictionary.pdf) for each table.

In total, we ended up with 1,7,749,767 rows, one row representing one admitted individual, and 27 variables. 23 of our variables are predictors. We engineered our response variable, wherein the DSMCRIT value was equal to “Opioid Abuse” or “Opioid Dependence”. A “1” in the DSMCRIT column represents an addiction to opiates, leaving the “0” to represent all non-additions. 3 of our columns are descriptive, and offer the FIPS, Date, and Case Number as unique identifiers.

# DATA PRE-PROCESSING

All data was collected from a variety of sources found in the Resources document, and are stored in separate CSV files, each of which was loaded individually to the local Spark cluster. All data consisted of numeric factors representing a value relative to the column. For example, the age column had values ranging from 2-12, representing age ranges from 12-14 through to 55 and over. All the factor variables were recoded twice, once to update the values to factors that were readable and understandable in visualizations, and then again, when all factors were treated with one-hot encoding to prepare it for the machine learning.

Additionally, since we intend on running the data through machine learning algorithms, we need to specify and correct individual columns for processing. This includes updating data types to numeric and factor where appropriate.

All these processes are found in the [raw code](https://github.com/sam-blumer/OpioidAddictionNN/blob/master/OpioidAddictionNN.R).

## MISSING DATA

Because all data was stored as factors relative to an individual, we could not simply impute the missing data without adding bias or erroneous data to the data set, so any row with a missing value needed in our analysis was simply removed.

# DATA EXPLORATION

With the data imported and cleaned, we can begin our data exploration process. Ultimately, we are looking at two different relationships. The first is identifying and exploring data points that are associated with being admitted to a rehabilitation center as a whole. This includes admission for reasons other than opioid use. Second is exploring how the characteristics of opioid users are similar. The data contains many demographic data points relative to each individual. These data points include age, income source, educational attainment, ethnic makeup, living arrangement, and more.

## *TIME SERIES EXPLORATION*

As noted earlier, there has been an extended period in which opioid use has increased. To confirm this, we can plot opioid overdoes death rates at both the state level and the aggregate US level.

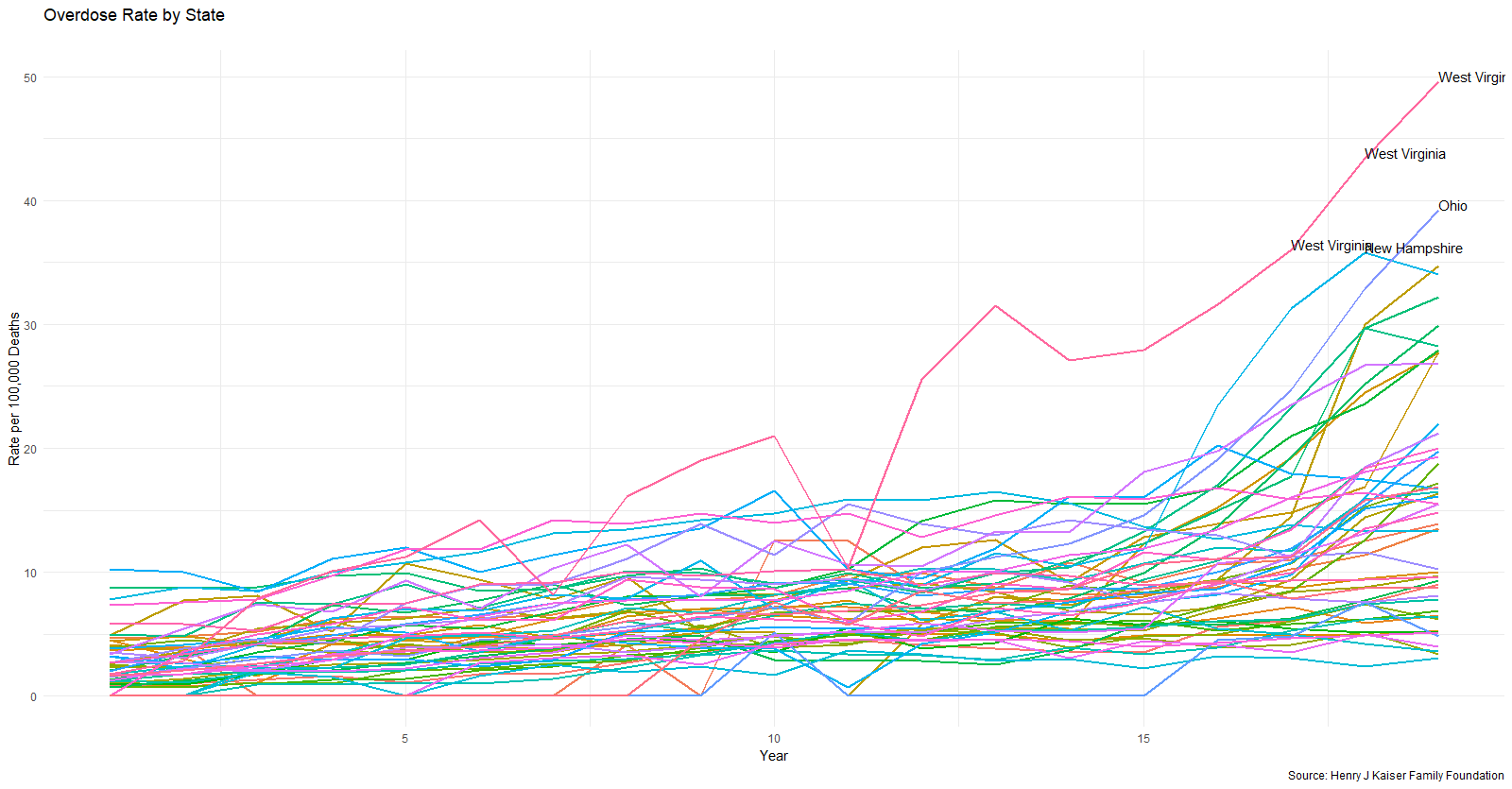


Figure 1

Figure 1 shows that nearly all states have had an increase in the opioid overdose death rate between 1999 and 2016. As mentioned previously, the states with the highest totals are those in the Midwest. We can see in Figure 3 below, the aggregate increase when looking at all states combined.

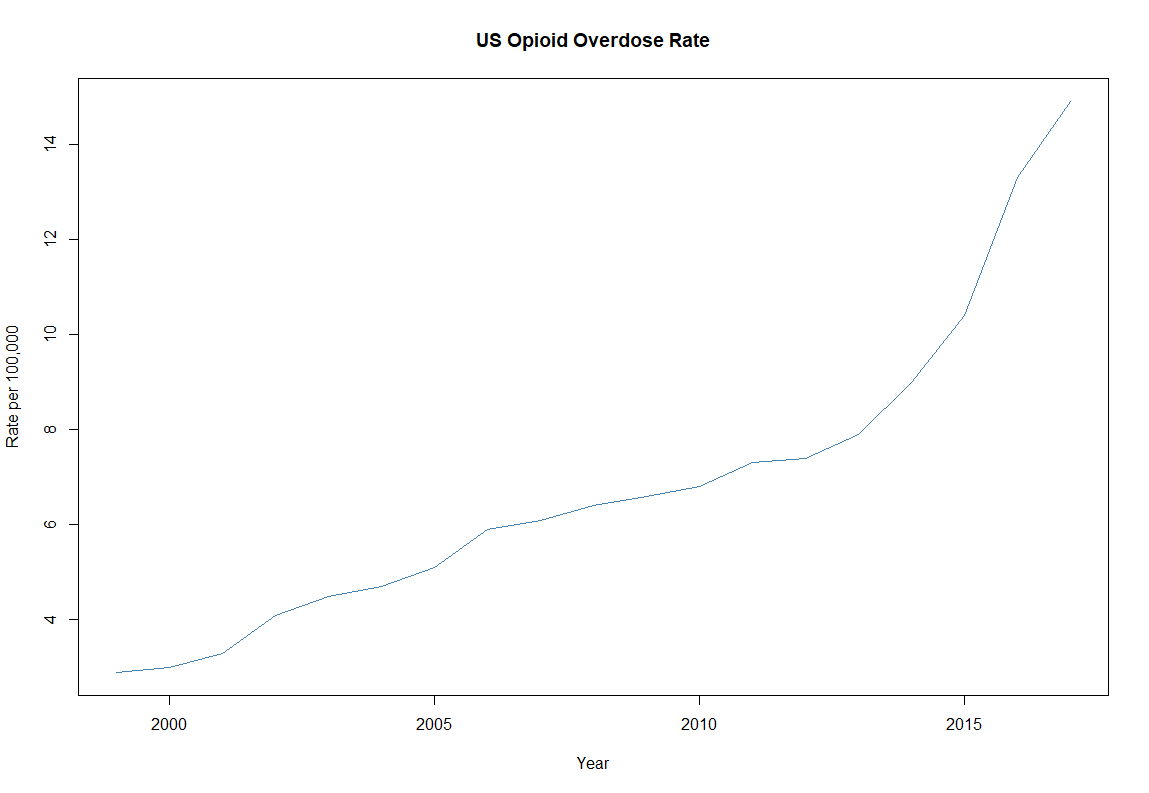


Figure 2

These charts confirm our research by showing that there has been a linear increase overall in opioid overdose death rates, giving reason to further our own studies an exploration. Because the dataset contains 1.7 million + rows related to all types of substance abuse, we are going to explore the data for both all substances, and then subset just the data that included opiates.

## *ALL ADMISSIONS EXPLORATION*

|  |  |
| --- | --- |
| OVERALL ADMISSION DATA | |
| **Metr**ic | Measure |
| Males | 66% |
| Female | 33% |
| White | 66 % |
| Black | 20% |
| HS or Less | 74% |
| Unemployed/Not in Work Force | 76% |
| Primary Income from Wages | 17.5% |
| Diagnosed Psychological Problem | 24% |
| Admitted Under Age 35 | 55% |
| First Substance Use Under Age 20 | 74% |

Figure 3

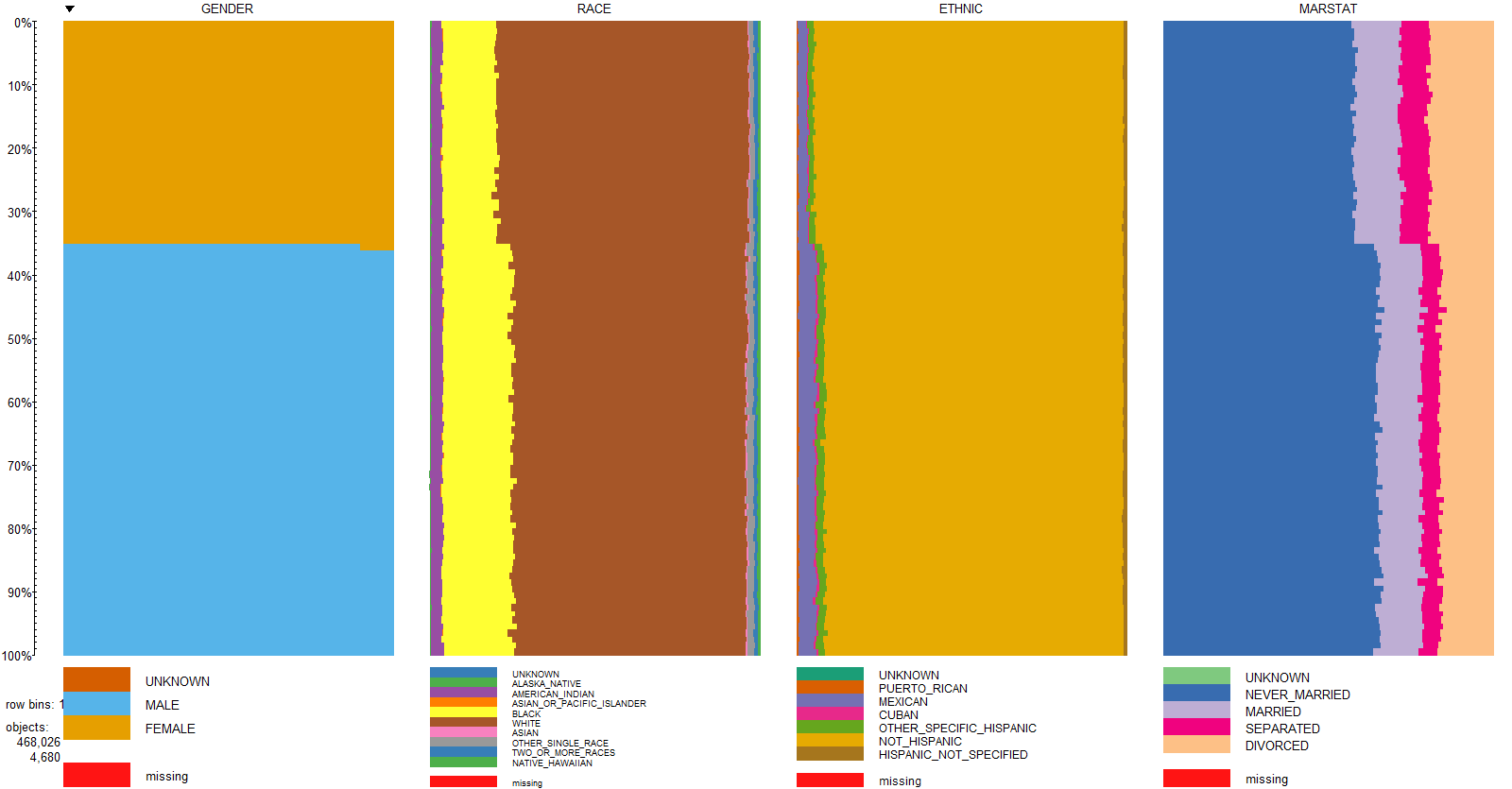


Figure 4

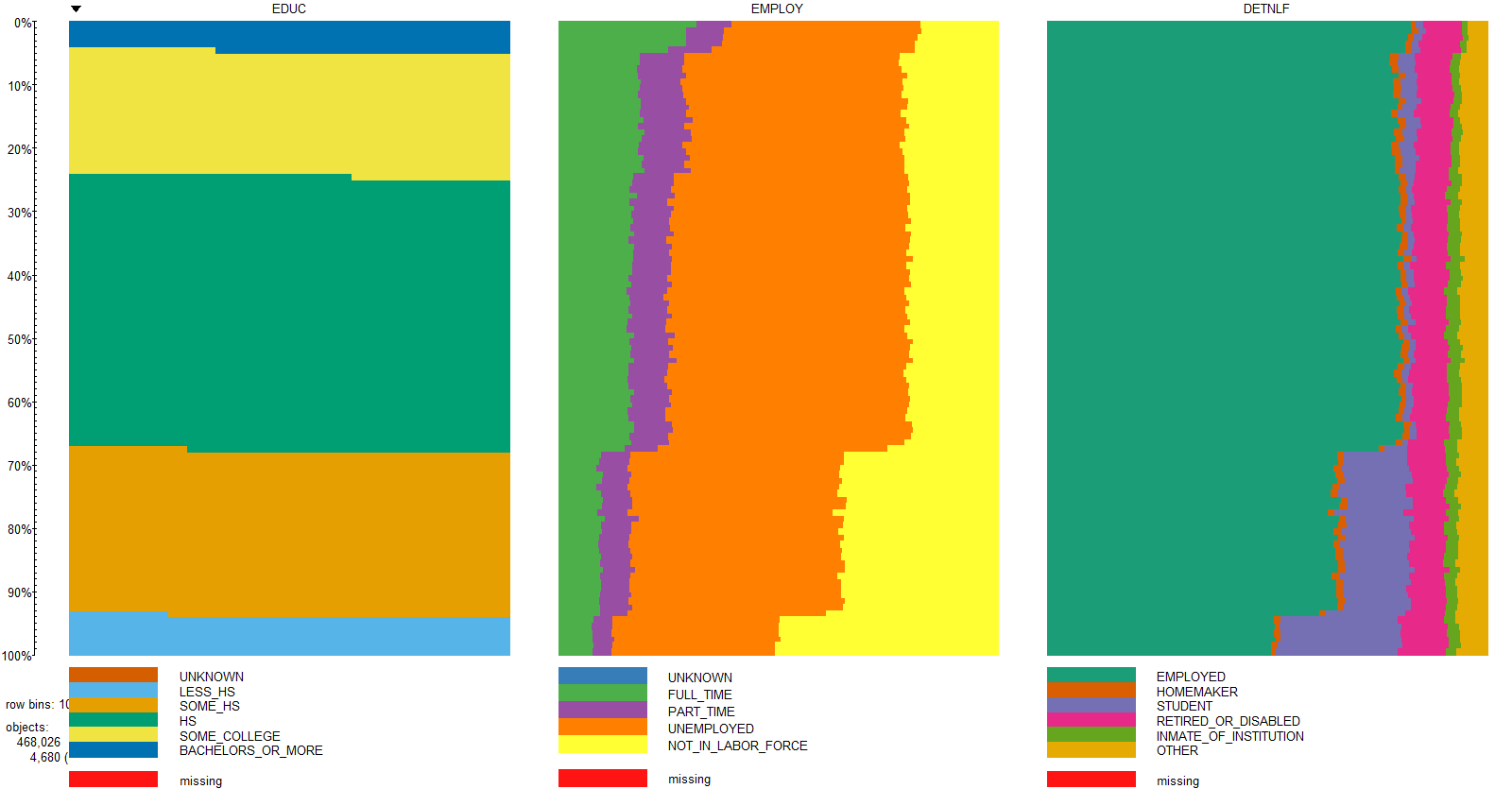


Figure 5

## *SUBSTANCE EXPLORATION*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Substance Name | Primary Substance | Secondary Substance | Tertiary Substance | Total Number | Percent of Patients |
| HEROIN | 251,519 | 40,319 | 14,783 | 306,621 | 17.55% |
| OTHER OPIATES AND SYNTHETICS | 151,328 | 79,390 | 32,451 | 263,169 | 15.07% |
| COCAINE OR CRACK | 109,927 | 167,902 | 73,447 | 351,276 | 20.11% |
| METHAMPHETAMINE | 71,172 | 34,243 | 18,023 | 123,438 | 7.07% |
| BENZODIAZEPINES | 16,403 | 52,995 | 31,586 | 100,984 | 5.78% |
| NONE | 14,251 | 667,713 | 149,801 | 831,765 | 47.62% |
| OTHER | 9,186 | 32,280 | 22,212 | 63,678 | 3.65% |
| OTHER AMPHETAMINES | 7,376 | 7,435 | 6,809 | 21,620 | 1.24% |
| PCP | 5,155 | 3,293 | 2,160 | 10,608 | 0.61% |
| NON-PRESCRIPTION METHADONE | 5,108 | 4,167 | 2,143 | 11,418 | 0.65% |
| OTHER NON-BARBITURATE SEDATIVES OR HYPNOTICS | 2,344 | 4,363 | 3,067 | 9,774 | 0.56% |
| OTHER HALLUCINOGENS | 1,886 | 3,426 | 5,199 | 10,511 | 0.60% |
| UNKNOWN | 1,104 | 3,676 | 207,386 | 212,166 | 12.15% |
| INHALANTS | 1,001 | 906 | 1,006 | 2,913 | 0.17% |
| OVER THE COUNTER MEDICATIONS | 968 | 1,319 | 1,097 | 3,384 | 0.19% |
| OTHER STIMULANTS | 909 | 1,238 | 1,170 | 3,317 | 0.19% |
| BARBITURATES | 673 | 783 | NA | 1,456 | 0.08% |
| OTHER NON-BENZODIAZEPINE TRANQUILIZERS | 336 | 764 | 717 | 1,817 | 0.10% |

Figure 6

Figure 6 demonstrates the substances found in a patient at admission. Data was recorded for up to three substances. For our study, we are including Heroin, Non-Prescription Methadone, and Other Opiates and Synthetics. The table above, then, shows that 33.25% of all admissions were opiate related.

## *OPIATE USER EXPLORATION*

To explore characteristics directly related to opiate users, we created a subset from our data the included Heroin, Non-Prescription Methadone, and Other Opiates and Synthetics and looked at the same demographics as in Figure 7.

|  |  |
| --- | --- |
| OVERALL ADMISSION DATA | |
| **Metr**ic | **Measure** |
| Males | 59% |
| Female | 41% |
| White | 77% |
| Black | 9% |
| HS or Less | 75% |
| Unemployed/Not in Work Force | 82% |
| Primary Income from Wages | 12% |
| Diagnosed Psychological Problem | 26% |
| Admitted Under Age 35 | 62% |
| First Substance Use Under Age 20 | 56% |

Figure 7

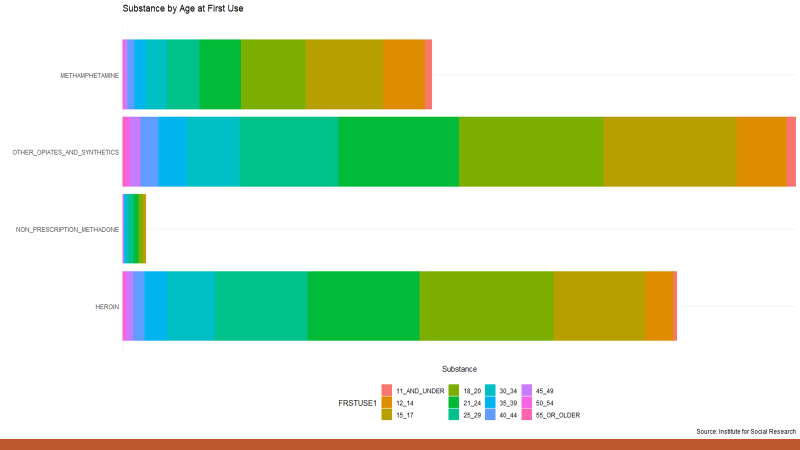


Figure 8

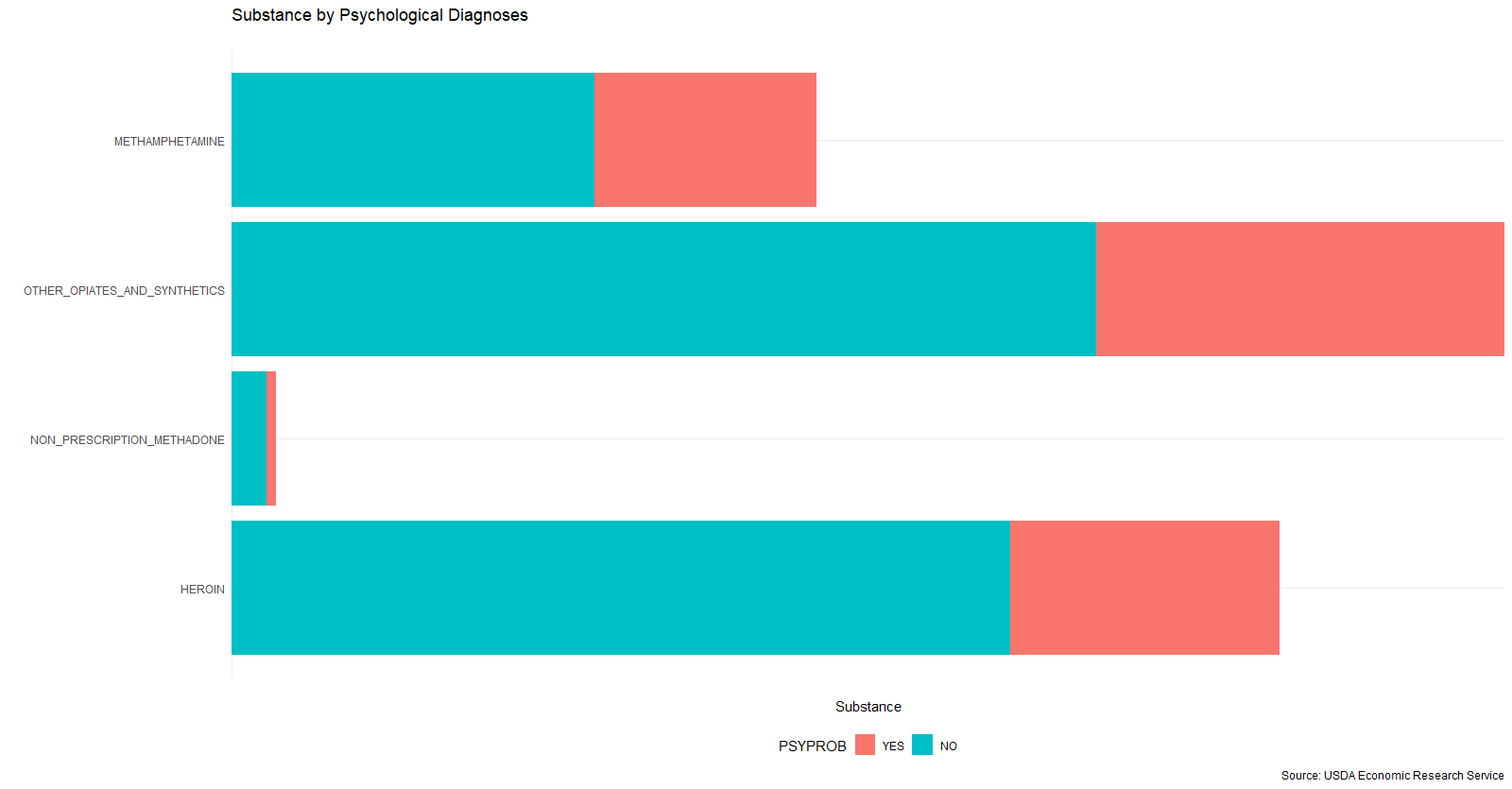


Figure 9

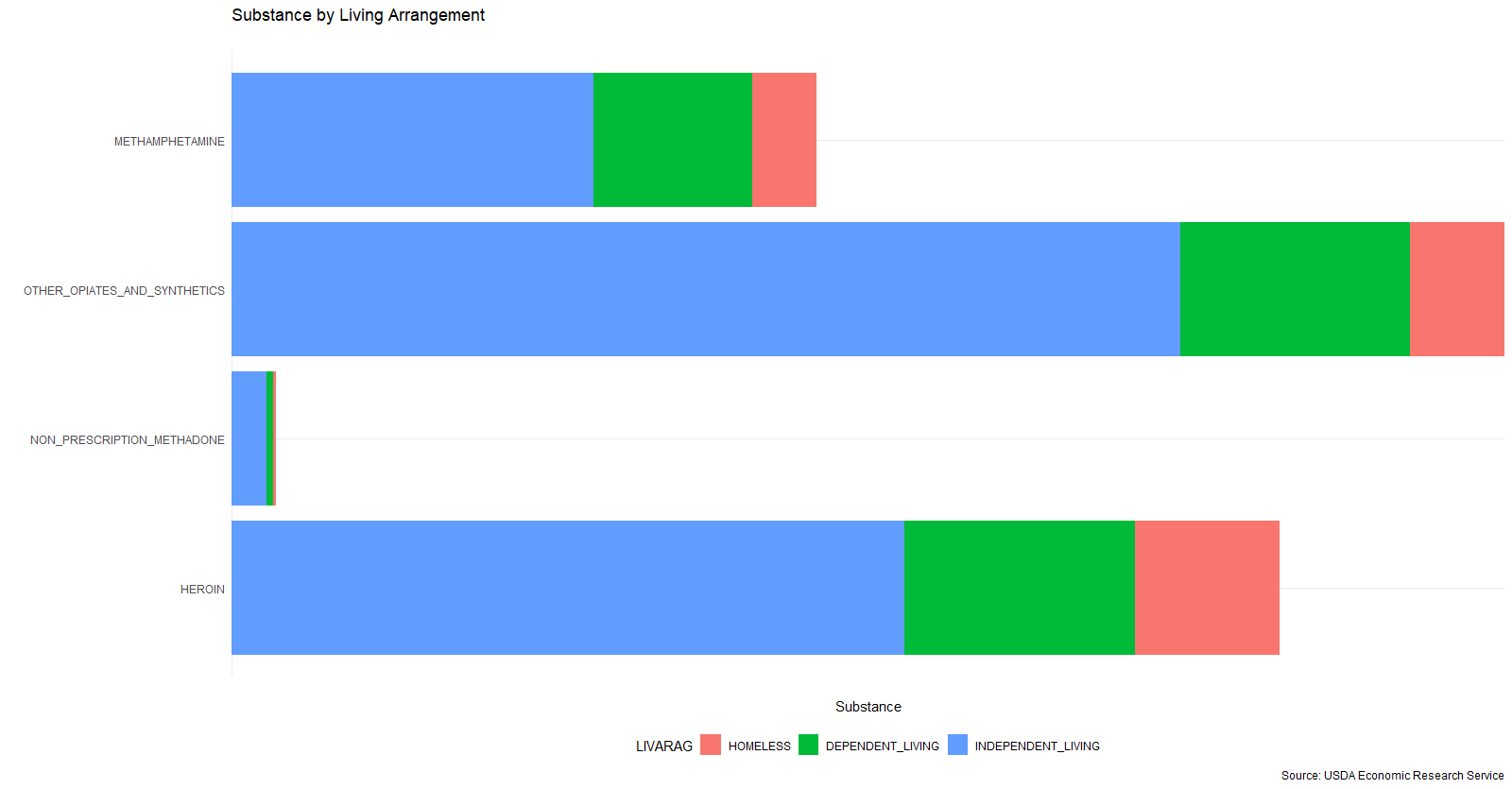


Figure 10

The data above tells us several things: While overall substance admission data tended to have twice the number of males, strictly opiate users tended to be equally female and male. They also tended to be white, unemployed, have just a high school education, live independently, and started using between ages 15 and 24.

# MODEL PREPARATION

For exploration, we used 29 variables that were all factors that represented some characteristic about the admitted individual. However, that data format cannot be fed to our Neural Nets, so we used one-hot encoding, ultimately ending up with 181 variables for 1.7 million+ observations. We also removed any confounding variables, for example, there was a but filed that represented the presence of Heroin, which would correlate with Substance 1, 2, or 3 being Heroin. With so many features, we decided to create a custom formula to use in both our feature engineering and model building:

Using all 180 predictors, our formula is as follows:

DSMCRIT ~ AGE\_12\_14 + AGE\_15\_17 + AGE\_18\_20 + AGE\_21\_24 + AGE\_25\_29 +

AGE\_30\_34 + AGE\_35\_39 + AGE\_40\_44 + AGE\_45\_49 + AGE\_50\_54 +

AGE\_55\_OR\_OLDER + GENDER\_UNKNOWN + GENDER\_MALE + GENDER\_FEMALE +

RACE\_UNKNOWN + RACE\_ALASKA\_NATIVE + RACE\_AMERICAN\_INDIAN +

RACE\_ASIAN\_OR\_PACIFIC\_ISLANDER + RACE\_BLACK + RACE\_WHITE +

RACE\_ASIAN + RACE\_OTHER\_SINGLE\_RACE + RACE\_TWO\_OR\_MORE\_RACES +

RACE\_NATIVE\_HAWAIIAN + ETHNIC\_UNKNOWN + ETHNIC\_PUERTO\_RICAN +

ETHNIC\_MEXICAN + ETHNIC\_CUBAN + ETHNIC\_OTHER\_SPECIFIC\_HISPANIC +

ETHNIC\_NOT\_HISPANIC + ETHNIC\_HISPANIC\_NOT\_SPECIFIED + MARSTAT\_UNKNOWN +

MARSTAT\_NEVER\_MARRIED + MARSTAT\_MARRIED + MARSTAT\_SEPARATED +

MARSTAT\_DIVORCED + EDUC\_UNKNOWN + EDUC\_LESS\_HS + EDUC\_SOME\_HS +

EDUC\_HS + EDUC\_SOME\_COLLEGE + EDUC\_BACHELORS\_OR\_MORE + EMPLOY\_UNKNOWN +

EMPLOY\_FULL\_TIME + EMPLOY\_PART\_TIME + EMPLOY\_UNEMPLOYED +

EMPLOY\_NOT\_IN\_LABOR\_FORCE + DETNLF\_EMPLOYED + DETNLF\_HOMEMAKER +

DETNLF\_STUDENT + DETNLF\_RETIRED\_OR\_DISABLED + DETNLF\_INMATE\_OF\_INSTITUTION +

DETNLF\_OTHER + VET\_UNKNOWN + VET\_VETERAN + VET\_NON\_VETERAN +

LIVARAG\_UNKNOWN + LIVARAG\_HOMELESS + LIVARAG\_DEPENDENT\_LIVING +

LIVARAG\_INDEPENDENT\_LIVING + PRIMINC\_NONE + PRIMINC\_WAGE\_OR\_SALARY +

PRIMINC\_PUBLIC\_ASSISTANCE + PRIMINC\_RETIREMENT\_OR\_DISABILITY +

PRIMINC\_OTHER + ARRESTS\_0 + ARRESTS\_1 + ARRESTS\_2\_OR\_MORE +

PSOURCE\_UNKNOWN + PSOURCE\_INDIVIDUAL + PSOURCE\_ALCOHOL\_OR\_DRUG\_CARE\_PROVIDER +

PSOURCE\_OTHER\_HEALTH\_PROVIDER + PSOURCE\_SCHOOL + PSOURCE\_EMPLOYER +

PSOURCE\_OTHER\_COMMUNITY\_REFERRAL + PSOURCE\_COURT\_OR\_CRIMINAL\_JUSTICE +

DETCRIM\_NONE + DETCRIM\_STATE\_OR\_FEDERAL\_COURT + DETCRIM\_PROBATION\_OR\_PAROLE +

DETCRIM\_DIVERSION\_PROGRAM + DETCRIM\_PRISON + DETCRIM\_DUI +

DETCRIM\_OTHER\_LEGAL\_ENTITY + NOPRIOR\_UNKNOWN + NOPRIOR\_0 +

NOPRIOR\_1 + NOPRIOR\_2 + NOPRIOR\_3 + NOPRIOR\_4 + NOPRIOR\_5\_OR\_MORE +

SUB1\_UNKNOWN + SUB1\_NONE + SUB1\_ALCOHOL + SUB1\_COCAINE\_OR\_CRACK +

SUB1\_MARIJUANA\_OR\_HASHISH + SUB1\_HEROIN + SUB1\_NON\_PRESCRIPTION\_METHADONE +

SUB1\_OTHER\_OPIATES\_AND\_SYNTHETICS + SUB1\_PCP + SUB1\_OTHER\_HALLUCINOGENS +

SUB1\_METHAMPHETAMINE + SUB1\_OTHER\_AMPHETAMINES + SUB1\_OTHER\_STIMULANTS +

SUB1\_BENZODIAZEPINES + SUB1\_OTHER\_NON\_BENZODIAZEPINE\_TRANQUILIZERS +

SUB1\_BARBITURATES + SUB1\_OTHER\_NON\_BARBITURATE\_SEDATIVES\_OR\_HYPNOTICS +

SUB1\_INHALANTS + SUB1\_OVER\_THE\_COUNTER\_MEDICATIONS + SUB1\_OTHER +

ROUTE1\_UNKNOWN + ROUTE1\_ORAL + ROUTE1\_SMOKING + ROUTE1\_INHALATION +

ROUTE1\_INJECTION + ROUTE1\_OTHER + FREQ1\_UNKNOWN + FREQ1\_NO\_USE\_IN\_PAST\_MONTH +

FREQ1\_1\_3\_TIMES\_IN\_THE\_PAST\_MONTH + FREQ1\_1\_2\_TIMES\_IN\_THE\_PAST\_WEEK +

FREQ1\_3\_6\_TIMES\_IN\_THE\_PAST\_WEEK + FREQ1\_DAILY + FRSTUSE1\_UNKNOWN +

FRSTUSE1\_11\_AND\_UNDER + FRSTUSE1\_12\_14 + FRSTUSE1\_15\_17 +

FRSTUSE1\_18\_20 + FRSTUSE1\_21\_24 + FRSTUSE1\_25\_29 + FRSTUSE1\_30\_34 +

FRSTUSE1\_35\_39 + FRSTUSE1\_40\_44 + FRSTUSE1\_45\_49 + FRSTUSE1\_50\_54 +

FRSTUSE1\_55\_OR\_OLDER + SUB2\_UNKNOWN + SUB2\_NONE + SUB2\_ALCOHOL +

SUB2\_COCAINE\_OR\_CRACK + SUB2\_MARIJUANA\_OR\_HASHISH + SUB2\_HEROIN +

SUB2\_NON\_PRESCRIPTION\_METHADONE + SUB2\_OTHER\_OPIATES\_AND\_SYNTHETICS +

SUB2\_PCP + SUB2\_OTHER\_HALLUCINOGENS + SUB2\_METHAMPHETAMINE +

SUB2\_OTHER\_AMPHETAMINES + SUB2\_OTHER\_STIMULANTS + SUB2\_BENZODIAZEPINES +

SUB2\_OTHER\_NON\_BENZODIAZEPINE\_TRANQUILIZERS + SUB2\_BARBITURATES +

SUB2\_OTHER\_NON\_BARBITURATE\_SEDATIVES\_OR\_HYPNOTICS + SUB2\_INHALANTS +

SUB2\_OVER\_THE\_COUNTER\_MEDICATIONS + SUB2\_OTHER + SUB3\_UNKNOWN +

SUB3\_NONE + SUB3\_ALCOHOL + SUB3\_COCAINE\_OR\_CRACK + SUB3\_MARIJUANA\_OR\_HASHISH +

SUB3\_HEROIN + SUB3\_NON\_PRESCRIPTION\_METHADONE + SUB3\_OTHER\_OPIATES\_AND\_SYNTHETICS +

SUB3\_PCP + SUB3\_OTHER\_HALLUCINOGENS + SUB3\_METHAMPHETAMINE +

SUB3\_OTHER\_AMPHETAMINES + SUB3\_OTHER\_STIMULANTS + SUB3\_BENZODIAZEPINES +

SUB3\_OTHER\_NON\_BENZODIAZEPINE\_TRANQUILIZERS + SUB3\_BARBITURATES +

SUB3\_OTHER\_NON\_BARBITURATE\_SEDATIVES\_OR\_HYPNOTICS + SUB3\_INHALANTS +

SUB3\_OVER\_THE\_COUNTER\_MEDICATIONS + SUB3\_OTHER + NUMSUBS\_0 +

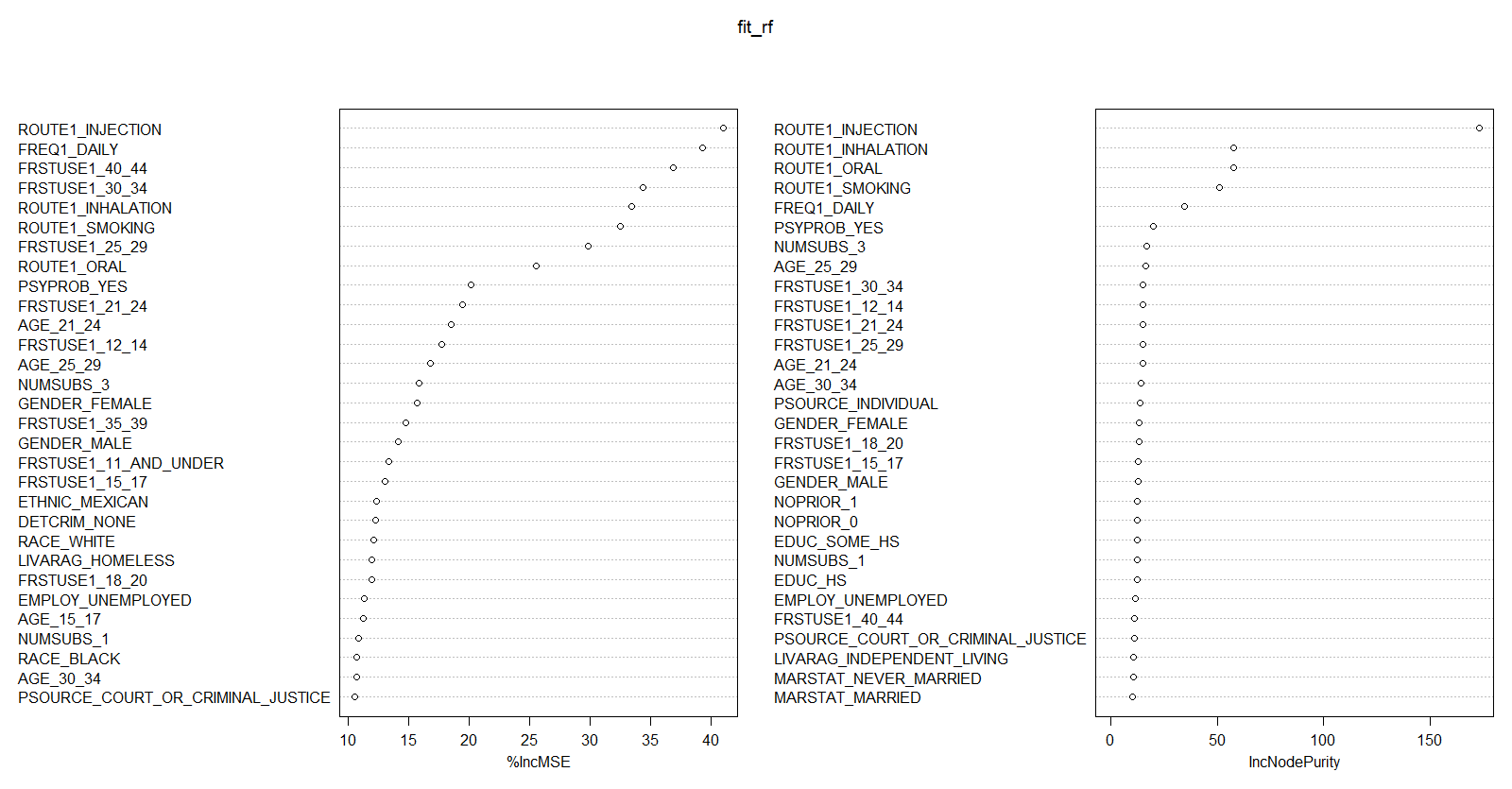
NUMSUBS\_1 + NUMSUBS\_2 + NUMSUBS\_3 + PSYPROB\_YES

Figure 11

# FEATURE ENGINEERING

# RANDOM FOREST FEAUTRE EXPLORATION

While Neural Net are meant to be tweaked by changing the architecture (hidden layers and nodes) and perform feature engineering and component analysis inherently, I wanted to see what features were most closely associated with being diagnosed with an opioid disorder. As mentioned previously, the dataset, at this point, contains 180 predictors, and it is likely that some of these impact our response more than others, and that some may not impact the response at all. I therefore used the Random Forest Variable Importance function to identify the variance of each variable on our dataset, the results for which are seen below.

Figure 12

From this analysis, we can derive some of our most important variables, including: the route the drug was administered (injection vs inhalation vs smoking), the frequency of use (daily, weekly), the age at which an individual started using (30-34, 40-44, 21-24), gender, and psychological problem diagnosis, for example, prove to be some of the biggest indicators, while factors like marriage status and living arrangement have less of an impact.

The RFVI model also uses a basic random forest to predict opiate addiction, the results of which are seen below.

Call:

randomForest(formula = as.formula(fla), data = data\_train, importance = TRUE, na.action = na.roughfix)

Type of random forest: regression

Number of trees: 500

No. of variables tried at each split: 60

Mean of squared residuals: 0.04566369

% Var explained: 69.62

Figure 13

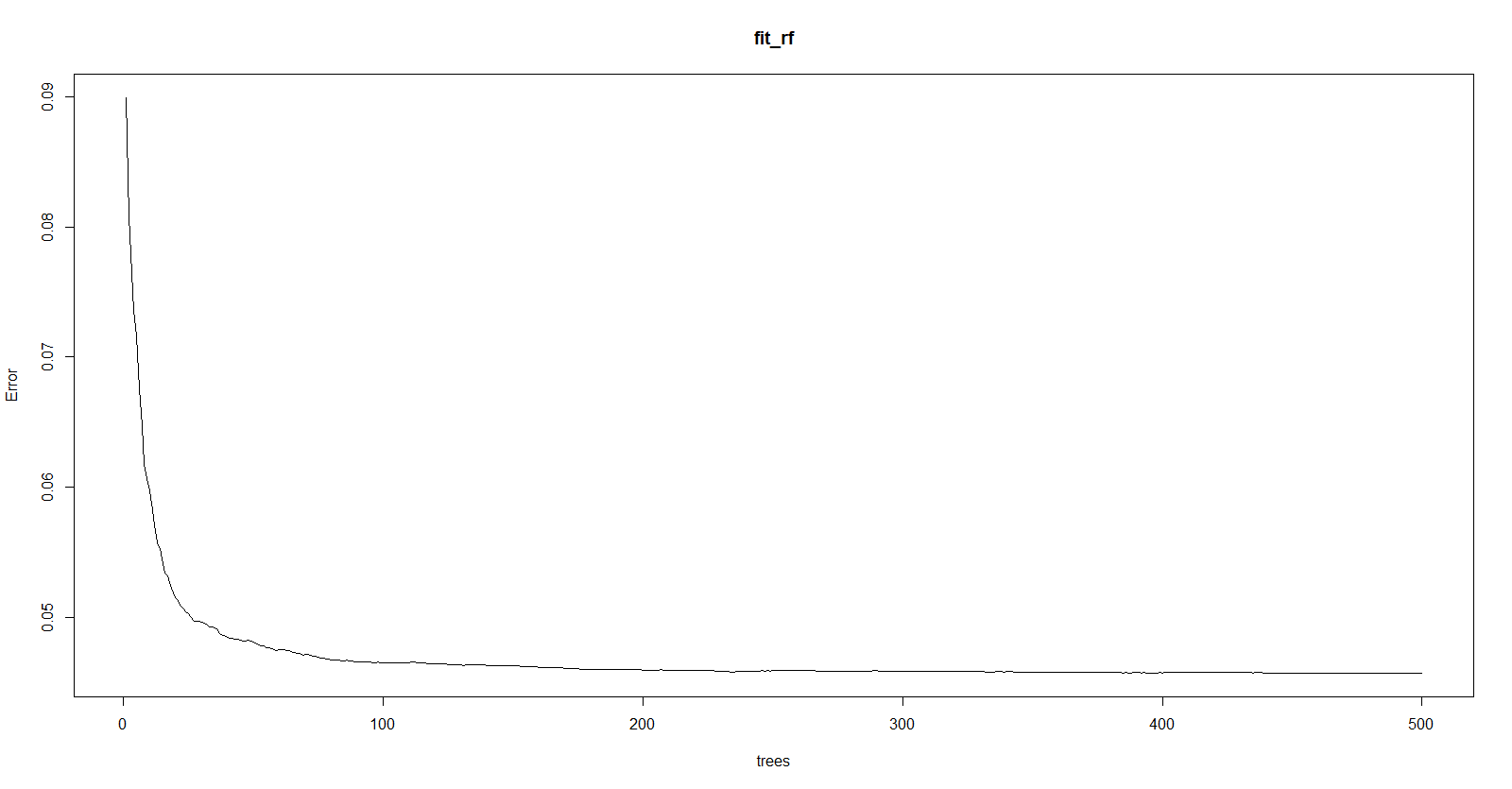


Figure 14

The RFVI used 500 trees with 60 variables split at each, however, we can see that the overall variance was low, and that the error rate remained mostly consistent after just 50 trees.

We split the data into a training and testing split of 80/20 and set the seed to 123 for reproducibility.

# 

Figure 15

# MODELS

We built four different models for comparisons sake. These included:

* + NN1: 2-Hidden Layers - Layer-1: 1-neuron, Layer-2: 1-neuron
  + NN2: 2-Hidden Layers - Layer-1: 2-neurons, Layer-2: 1-neuron
  + NN3: 2-Hidden Layers, Layer-1: 2-neurons, Layer-2: 2-neurons
  + NN4: 2-Hidden Layers, Layer-1: 1-neuron, Layer-2: 2-neurons

To make sure that our models were more likely to be true than not, and to compare performance, we sue the Bayesian Information Criterion and the Alkaline Information Criterion, alongside overall error. For these reference points, a lower value is better, hence Figure 16 below indicates that our most accurate model was NN3, which had two hidden layers, both with 2 neurons.

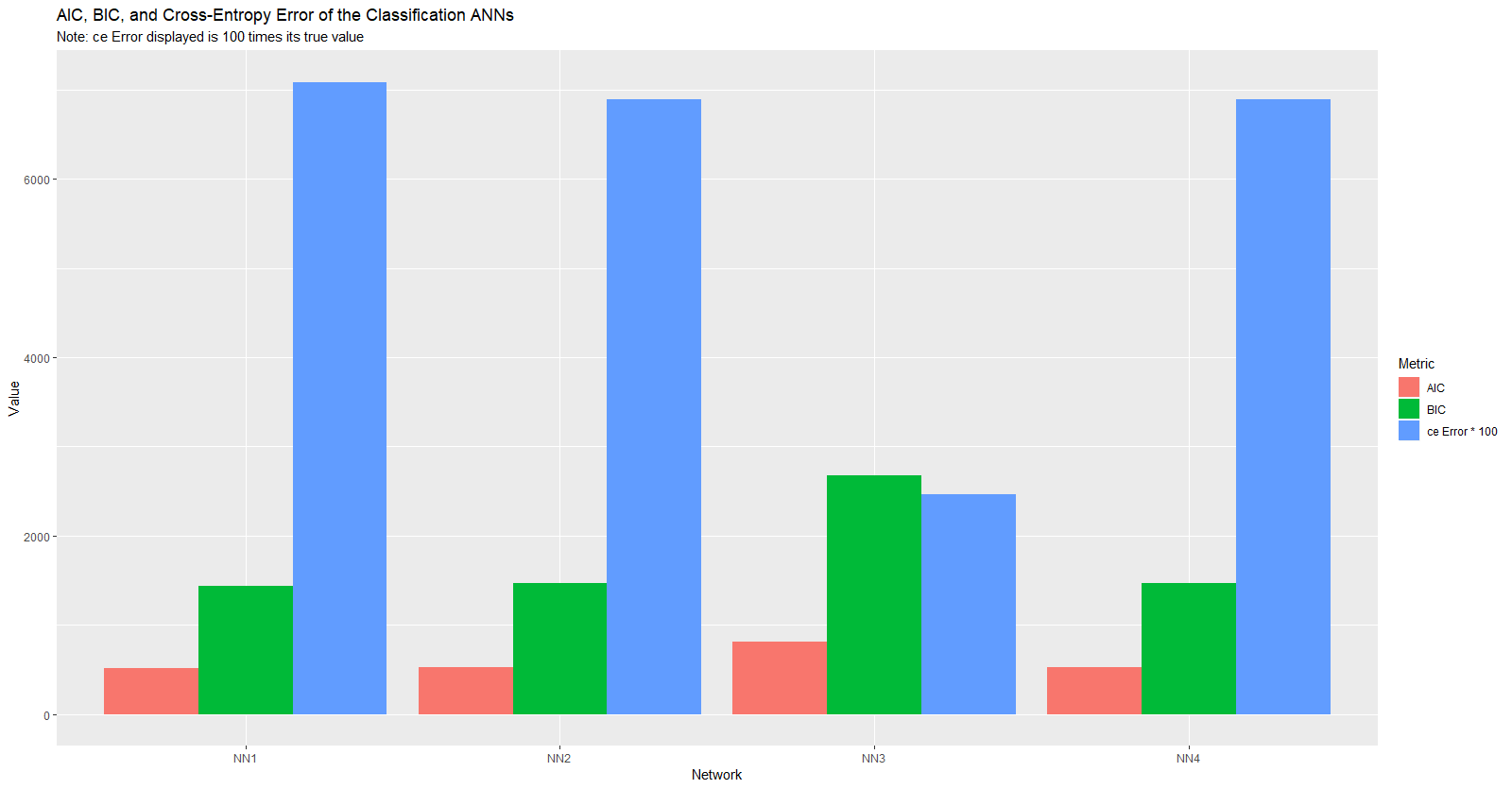


Figure 16

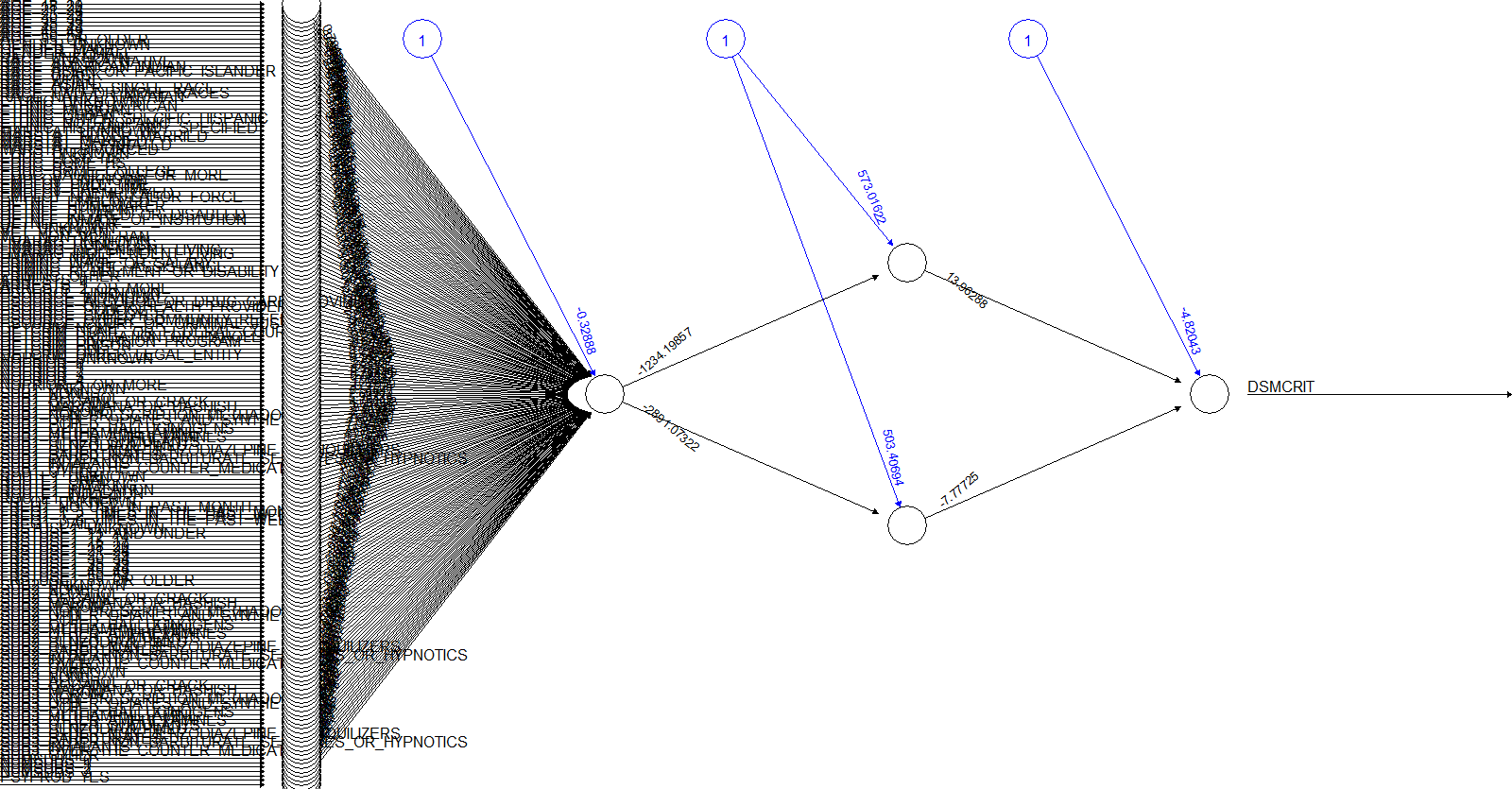


Figure 17

Since NN3 looked to be the most accurate, we created a confusion matric to evaluate the performance. As we can see in Figure 18, the measures were as follows:

* **Accuracy** = (1520 + 358)/2000 = **.94**
* **Recall** = 1520 / (1520 +38) = **.98**
* **Precision** = 1520 / (1520 + 84) = **.95**

|  |  |  |
| --- | --- | --- |
|  | Predictions | |
| **Actuals** | 0 | 1 |
| 0 | 1,520 | 84 |
| 1 | 38 | 358 |

Figure 18

# Conclusion

Overall, we were able to identify that with 180 predictor variables, we were able to most accurately predict food deserts with a 94% accuracy rate. When using this neural net architecture to predict potential opiate addiction based off 2012 data, we found that our results were consistent with initial exploratory findings, including:

* White
* Male
* Unemployed
* Route of Drug Administration
* Age at First Use
* Other Substance Use