Final Project

Team 20

```
library(knitr)
library(tidyverse)
library(broom)
library(htmltools)
library(caret)
library(tidymodels)
library(schrute)
library(lubridate)
library(knitr)
library(openintro)
library(ROSE)
library(dplyr)
```

1. Introduction

The research question we will focus on is the characteristics of an individual using hard drugs. In general, many people don't understand why or how other people become addicted to drugs. They may mistakenly think that those who use drugs lack moral principles or willpower and that they could stop their drug use simply by choosing to. In reality, drug addiction is a complex disease, and quitting usually takes more than good intentions or a strong will.

A significant portion of the global population uses hard drugs. There are many dangers associated with the use of hard drugs, and over the years, accessibility to them has been increasing among various populations. The goal of our research is to understand the underlying factors contributing to addiction to hard drugs. We aim to focus on different populations to combat the phenomenon in a targeted and widespread manner. One of the reasons why understanding the characteristics of individuals using hard drugs can be challenging is the complex nature of drug addiction itself. Drug addiction is a multifaceted issue influenced by various biological, psychological, and social factors. These factors interact in unique ways for each individual, making it difficult to generalize or predict patterns of drug use and addiction.

Our approach is to explore the information, while cleaning it and finding strong connections between different variables using the Cramer's V correlation method, which is known for categorical variables. After understanding the strong correlations, we will perform column factorization since all our variables represent categorical variables. We will create a logistic regression model using the relevant column to predict whether a person is registered in addiction treatment institutions due to a diagnosis of hard drug addiction. We will improve the model by filtering out columns that are not significant in the model and reach the relevant columns for our model. Based on the results, we will obtain characteristics of a person addicted to hard drugs.

2. Data

The data set contains information on the admission of individuals aged 12 years or older, including their demographic characteristics such as age, gender, race, ethnicity, employment status, and substance use

patterns. This data was collected in 2020 in the United States. Specifically, we will focus on variables related to the person's characteristics, such as their gender, age at admission, marital status, education, employment status, living arrangements, and source of income.

The dataset contains a column named DSMCRIT representing the initial diagnosis of a person entering a drug addiction treatment facility. There are several levels of diagnoses, including diagnoses related to mental conditions such as depression, anxiety, etc., as well as diagnoses related to addiction to different types of drugs, including hard drugs.

3. Methods and results

In the dataset we are working on, there is a column called DSMCRIT which, as mentioned earlier, represents the initial diagnosis of a person entering a drug addiction treatment facility. We chose to represent this column in a binary form as HardDrugs, where individuals diagnosed as addicted to hard drugs are represented as 1, and the rest are represented as 0.

There were around 400,000 missing values in DSMCRIT, so we completed these values using the columns sub1, sub2, and sub3, which represent the primary, secondary, and tertiary substances declared at registration.

We explored several options to fill in the missing values based on the correspondence between the binary column HardDrugs and the binary columns we created for sub1, sub2, and sub3, indicating whether a hard drug was used or not.

An analysis of the three binary columns showed a correlation between using two or more hard drugs and being diagnosed as addicted to hard drugs. Therefore, we decided that the use of two or more hard drugs would be classified as 1 in the HardDrugs column, while the rest would be classified as 0.

The accuracy of the match value in HardDrugs for records that have a diagnosis in DSMCRIT

```
## Percentage of identical rows: 64.4686 %
```

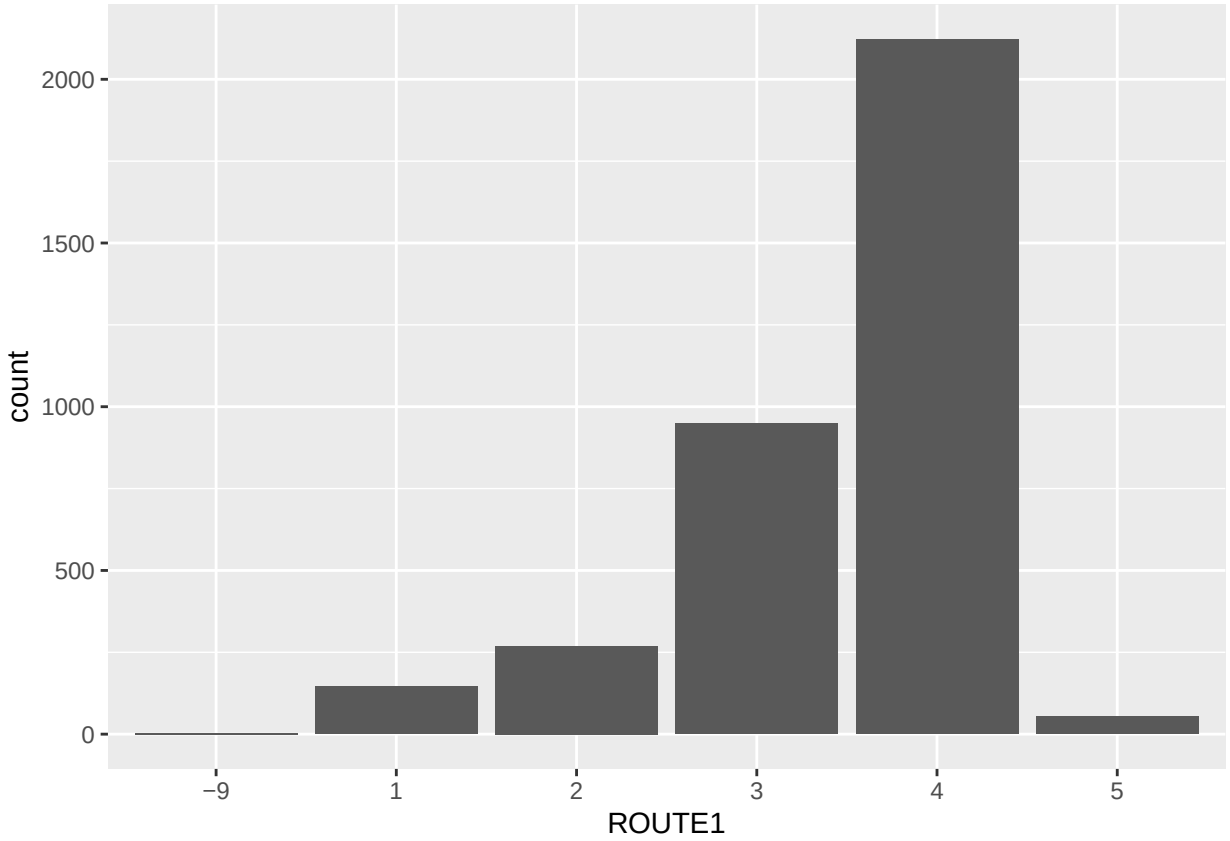
```
##   Value Correct_Percentage Incorrect_Percentage
## 1     0           62.15625             37.84375
## 2     1           73.17071             26.82929
```

We built a correlation matrix based on Cramer for the explanatory variables, and we filtered out variables that have a high correlation between them in order not to bias the model.

After completing the values, we took a 20% sample of the data and divided it into train and test sets to build a logistic regression model. We constructed the model iterativly and filtered the columns based on their p-values until we obtained a model with good metrics, using only the most significant columns.

```
##        Test     Value
## 1  Accuracy 0.8414104
## 2 Precision 0.8748100
## 3    Recall 0.8676776
## 4 F-measure 0.8712292
```

After conducting the prediction, we took the top 5% of records with the highest probability to be classified as 1 - diagnosed as addicted to hard drugs. We then performed the analysis and drew conclusions about the characteristics of the population of individuals addicted to hard drugs entering the rehabilitation institution.



Using these graphs, we have drawn conclusions about each variable we chose to analyze in the table. Here are the conclusions we reached:

Personal characteristics:

Marital status: Single.

Common age range: 30-34.

Ethnicity: White, referring to a person of European, Middle Eastern, or North African origin.

Living alone.

Education: grade 12 (or GED).

Gender: No significant difference between males and females.

Originating from New Jersey and Michigan - urban areas.

Socio-economic status:

Unemployed - without a source of income.

Having government health insurance, therefore, treatment is funded by the government.

Addiction characteristics:

Most individuals went through 5 or more treatment cycles in addiction facilities.

Half of the individuals addicted to hard drugs have a mental disorder, a higher percentage compared to other addictions.

Most individuals voluntarily sought treatment.

The most common method of drug use is daily injection.

Most individuals started using drugs at the age of 30. It is noteworthy that the common age range for drug addicts is 30-34, suggesting rapid deterioration.

The vast majority do not require alcohol and marijuana.

We expected higher percentages of cocaine/crack users, but the majority of addicts primarily use heroin.

4. Limitations and Future Work

One of the limitations we found is that the data contains information regarding addicts attending rehabilitation institutions but does not include general information about addicts. Additionally, we were limited in terms of the amount of data the model could process. We believe we would have obtained slightly different results if we had processed all the records in the data.

Another limitation is the lack of responsiveness and guidance from the researcher regarding the data. Despite our extensive work and investment, we still feel there is a professional gap.

If we had more time, we would have included data collected in previous years in the model to examine the differences between years and different addiction characteristics.

Appendix

Link to a code repository: https://github.com/nofarselouk/Advanced-programming/blob/main/project.Rmd

Data README

Source code

```
library(knitr)
library(tidyverse)
library(broom)
library(htmltools)
library(caret)
library(tidymodels)
library(schrute)
library(lubridate)
library(knitr)
library(openintro)
library(ROSE)
library(dplyr)
opts_chunk$set(echo=FALSE) # hide source code in the document
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
drugsa <- read.csv("C:/Users/nofar/Desktop/TEDSA_PUF_2020.csv")
#Fills the "HardDrugs" column based on the presence of frequent drugs

fill_HardDrugs <- function(data) {
    for (i in 1:nrow(data)) {
      if (data$DSMCRIT[i] == -9) {
        s <- sum(c(data$sub1_bin[i], data$sub2_bin[i], data$sub3_bin[i]))

        if (s >= 2) {
```

```
          data$HardDrugs[i] <- 1}
        else{
          data$HardDrugs[i] <- 0}
    }
  }

  # Return the updated data
  return(data)
}
#' Performs data manipulation and calculates metrics related to the "HardDrugs" column
#'
#Subset the dataset based on ASNCRIT values
subset <- drugsa[drugsa$DSMCRIT != -9, ]
#Assign values to "HardDrugs" based on DSMCRIT values
subset$HardDrugs[subset$DSMCRIT %in% c(1, 2, 3, 4, 7, 9, 10, 14, 15, 16, 17, 18, 19)] <- 0
subset$HardDrugs[subset$DSMCRIT %in% c(5, 6, 8, 11, 12, 13)] <- 1

hard_drug_codes <- c(3, 5, 7, 8, 9, 10, 12, 13,17,19)
#Update "sub1_bin", "sub2_bin", and "sub3_bin" columns
subset$sub1_bin[subset$SUB1 %in% c(1, 2, 4, 6, 11, 14, 15, 16, 18, -9)] <- 0
subset$sub1_bin[subset$SUB1 %in% hard_drug_codes] <- 1
subset$sub2_bin[subset$SUB2 %in% c(1, 2, 4, 6, 11, 14, 15, 16, 18, -9)] <- 0
subset$sub2_bin[subset$SUB2 %in% hard_drug_codes] <- 1
subset$sub3_bin[subset$SUB3 %in% c(1, 2, 4, 6, 11, 14, 15, 16, 18, -9)] <- 0
subset$sub3_bin[subset$SUB3 %in% hard_drug_codes] <- 1

# Calculate the sum of the values in the specified columns
sum_values <- rowSums(subset[, c("sub1_bin", "sub2_bin", "sub3_bin")])

# Create the new binary column
subset$check_hashlama <- ifelse(sum_values >= 2, 1, 0)

# Calculate the percentage of rows where the values in the columns are identical
identical_rows <- sum(subset$HardDrugs == subset$check_hashlama)
percentage_identical <- (identical_rows / nrow(subset)) * 100

# Print the result
cat("Percentage of identical rows:", percentage_identical, "%\n")


# Calculate the number of correct and incorrect values for each value (0 and 1)
correct_0 <- sum(subset$HardDrugs[subset$check_hashlama == 0] == 0)
correct_1 <- sum(subset$HardDrugs[subset$check_hashlama == 1] == 1)
incorrect_0 <- sum(subset$HardDrugs[subset$check_hashlama == 0] == 1)
incorrect_1 <- sum(subset$HardDrugs[subset$check_hashlama == 1] == 0)

# Calculate the percentages
percentage_correct_0 <- correct_0 / sum(subset$check_hashlama == 0) * 100
percentage_correct_1 <- correct_1 / sum(subset$check_hashlama == 1) * 100
percentage_incorrect_0 <- incorrect_0 / sum(subset$check_hashlama == 0) * 100
percentage_incorrect_1 <- incorrect_1 / sum(subset$check_hashlama == 1) * 100

# Create a table
```

```r
table_data <- data.frame(
  Value = c(0, 1),
  Correct_Percentage = c(percentage_correct_0, percentage_correct_1),
  Incorrect_Percentage = c(percentage_incorrect_0, percentage_incorrect_1)
)

# Print the table
print(table_data)
fdata <- drugsa[, !(names(drugsa) %in% c("CASEID", "ADMYR", "CBSA2010"))]

# Set the values to 1 where the DSMCRIT values match the specified conditions
fdata$HardDrugs[fdata$DSMCRIT %in% c(1, 2, 3, 4, 7, 9, 10, 14, 15, 16, 17, 18, 19)] <- 0
fdata$HardDrugs[fdata$DSMCRIT %in% c(5, 6, 8, 11, 12, 13)] <- 1


hard_drug_codes <- c(3, 5, 7, 8, 9, 10, 12, 13, 17, 19)

fdata$sub1_bin[fdata$SUB1 %in% c(1, 2, 4, 6, 11, 14, 15, 16, 18, -9) & fdata$DSMCRIT == -9] <- 0
fdata$sub1_bin[fdata$SUB1 %in% hard_drug_codes & fdata$DSMCRIT == -9] <- 1
fdata$sub2_bin[fdata$SUB2 %in% c(1, 2, 4, 6, 11, 14, 15, 16, 18, -9) & fdata$DSMCRIT == -9] <- 0
fdata$sub2_bin[fdata$SUB2 %in% hard_drug_codes & fdata$DSMCRIT == -9] <- 1
fdata$sub3_bin[fdata$SUB3 %in% c(1, 2, 4, 6, 11, 14, 15, 16, 18, -9) & fdata$DSMCRIT == -9] <- 0
fdata$sub3_bin[fdata$SUB3 %in% hard_drug_codes & fdata$DSMCRIT == -9] <- 1

# Calculate the sum of the values in the specified columns
sum_values <- rowSums(fdata[, c("sub1_bin", "sub2_bin", "sub3_bin")])

# Create the new binary column
fdata <- fill_HardDrugs(fdata)

library(vcd)

# Get column names of drugs1 dataset
column_names <- colnames(fdata)

correlation <- vector("numeric", length(column_names))
for (i in seq_along(column_names)) {
  contingency_table <- table(fdata$HardDrugs, fdata[, i])
  correlation[i] <- assocstats(contingency_table)$cramer
}

# Sort correlations in descending order
sorted_correlation <- sort(correlation, decreasing = TRUE)

# Print correlations with column names
#for (i in seq_along(sorted_correlation)) {
 # print(paste0("Correlation with ", column_names[i], ": ", sorted_correlation[i]))
#}
library(vcd)

# Get column names of the dataset
column_names <- colnames(fdata)
```

```r
# Create an empty matrix to store correlations
correlation_matrix <- matrix(0, ncol = length(column_names), nrow = length(column_names))
rownames(correlation_matrix) <- column_names
colnames(correlation_matrix) <- column_names

# Calculate Cramer's V correlation for each column combination
for (i in seq_along(column_names)) {
  for (j in seq_along(column_names)) {
    contingency_table <- table(fdata[, i], fdata[, j])
    correlation_matrix[i, j] <- assocstats(contingency_table)$cramer
  }
}

selected_columns <- unlist(column_names[sorted_correlation >= 0.1])
selected_columns <- c(selected_columns, "HardDrugs")
selected_columns <- selected_columns[selected_columns != "DSMCRIT"]

unselected_columns <- column_names[sorted_correlation < 0.1]
unselected_columns <- c(unselected_columns, "DSMCRIT")
unselected_columns <- unselected_columns[unselected_columns != "HardDrugs"]
unselected_columns <- unlist(unselected_columns)

# Set the seed for reproducibility
set.seed(123)

# Take a random sample of five percent of the dataset
sampled_data <- fdata %>%
  sample_frac(0.2)


fdata_new <- select(sampled_data, selected_columns)
set.seed(666667)  # Set a seed for reproducibility
datasplit <- initial_split(fdata_new)
trainData <- training(datasplit)  # Training data
testData <- testing(datasplit)   # Testing/validation data
trainData <- trainData %>%
  mutate_all(factor)
testData <- testData %>%
  mutate_all(factor)

drugs_mod <- logistic_reg() %>%
  set_engine("glm")

drugs_rec <- recipe(HardDrugs ~ ., data = trainData) %>%
  step_rm(SUB1, SUB2, SUB3, DETNLF, DETCRIM, ROUTE3, FREQ3, FRSTUSE3, PREG, DAYWAIT, GENDER, METHFLG, SI
  step_dummy(all_nominal(), -all_outcomes()) %>%
  step_zv(all_predictors())
  #step_backward()

drugs_wflow <- workflow() %>%
  add_model(drugs_mod) %>%
  add_recipe(drugs_rec)
```

```r
logreg_fit <- drugs_wflow %>%
  fit(data = trainData)


results <- testData %>%
  bind_cols(logreg_fit %>% predict(new_data = testData, type = "prob"))%>%
              bind_cols(logreg_fit %>% predict(new_data = testData, type = "class"))

#conf_mat(data = results, truth = HardDrugs, estimate = .pred_class)
library(yardstick)
accuracy_val <- accuracy(data = results, truth = HardDrugs, estimate = .pred_class)
precision_val <- precision(data = results, truth = HardDrugs, estimate = .pred_class)
recall_val <- recall(data = results, truth = HardDrugs, estimate = .pred_class)
f_meas_val <- f_meas(data = results, truth = HardDrugs, estimate = .pred_class)


#tidy(logreg_fit)

# Extract the value from accuracy_val tibble
accuracy_value <- accuracy_val$.estimate

# Extract the value from precision_val tibble
precision_value <- precision_val$.estimate

# Extract the value from recall_val tibble
recall_value <- recall_val$.estimate

# Extract the value from f_meas_val tibble
f_measure_value <- f_meas_val$.estimate

# Create a data frame to store the values
table_data <- data.frame(Test = c("Accuracy", "Precision", "Recall", "F-measure"),
                         Value = c(accuracy_value, precision_value, recall_value, f_measure_value))

# Print the table
print(table_data)
# Calculate the number of rows to extract
num_rows <- ceiling(0.05 * nrow(results))

# Sort the results by descending order of ".pred_1" values
results_sorted <- results[order(-results$.pred_1), ]

# Extract the top 5% rows with highest ".pred_1" values
top_rows <- results_sorted[1:num_rows, ]

# Print the top rows
#print(top_rows)
columns <- colnames(top_rows)
columns <- columns[1:(length(columns) - 4)]
#create_bar_graph <- function(data, columns) {
#  for (column in columns) {
 #   print(ggplot(data, aes(x = .data[[column]])) + geom_bar())}
#}
```

```
#create_bar_graph(top_rows, columns)
print(ggplot(top_rows,aes(x = ROUTE1)) + geom_bar())
#cat(readLines('../data/README.md'), sep = '\n')
```