Caroline Crumrine, Caleigh Stivers, Cooper Helms
CSC-272 Final Project Report
April 22, 2024
Dr. Treu

Unraveling Connections: Personality Traits, Drug Usage Frequency, and Predictive Modeling for Legal versus Illegal Drug Consumption and Gateway Drug Identification

Introduction

As drug usage and addiction has become a worldwide global health concern, researchers want to determine the underlying factors that lead to substance consumption and abuse. Recently, underlying factors of interest include socio-demographic and personality characteristics. Therefore, exploration and analysis of comprehensive datasets relating to an individual's background, physiological traits, and associated drug usage serve as instrumental in discovering patterns in substance consumption and addiction behaviors.

With the use of various data pre-processing and mining techniques on the UC Irvine Machine Learning Repository's Drug Consumption Quantified dataset, we sought to explore the relationships between demographic traits, psychological characteristics, and the propensity for drug consumption. Such insights could lead to targeted interventions, preventative measures, and policy initiatives to reduce the prevalence and impact of drug addiction.

Dataset Description

The Drug Consumption Quantified dataset from the UCI Machine Learning Repository is a collection of data from 1885 adult participants from surveys and self-reported responses regarding their usage of various legal and illegal substances. The dataset includes attributes such as personality traits measured from psychological assessments, demographic details such as age, race, gender, education, and country of residence, and responses to a questionnaire about the frequency of drug consumption for eighteen different legal and illegal substances. The psychological assessments used to measure each subject's personality include the NEO-FFI-R, which calculates an individual's level of neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness, the BIS-11, a scale that measures impulsivity, and the Imp-SS, a tool used to calculate one's sensation seeking personality trait. Each assessment returns a numerical score with higher values representing high levels of a personality trait in a participant. The drugs included in the dataset range from legal substances such as chocolate and caffeine to illicit drugs such as crack and heroin. Interestingly, the dataset also included a fictitious substance, semeron, used to identify participants who provided inaccurate responses regarding personal drug usage. Each test subject selected one of seven classes describing their drug usage for each substance: never used, used more than 10 years ago, used in the last ten years, used in the last year, used in the last month, used in the last day. This dataset serves as a

resource for researchers and analysts aiming to understand personality and demographic factors/patterns influencing drug consumption behaviors.

Data Attributes

Attribute	**Description**
individual_id: *Numeric*†	ID number of the participant
age: *Nominal*†	Age of the participant. Its values fall into the following categories: ● 18-24 ● 25-34 ● 35-44 ● 45-54 ● 55-64 ● 65-
gender: *Nominal*†	Gender of the participant. Its values fall into the following categories: ● Female ● Male
education: *Nominal*†	Highest level of education completed by the participant: Its values fall into the following categories: ● Before_16: Dropped out before age 16 ● At_16: Dropped out at age 16 ● At_17: Dropped out at age 17 ● At_18: Dropped out at age 18 ● Some_College: Some college, no certificate/degree ● Certificate: Received trade school certificate ● Undergrad: Received undergraduate degree ● Masters: Received masters degree ● Doctorate: Received doctorate degree
country: *Nominal*†	Participant's current country of residence. Its values fall into the following categories: ● Australia ● Canada ● New Zealand ● Ireland ● UK ● USA ● Other

ethnicity: *Nominal*†	Participant's ethnicity. Its values fall into the following categories. ● Asian ● Black ● Mixed_BA: Mixed Black/Asian ● Mixed_WA: Mixed White/Asian ● Mixed_WB: Mixed White/Black ● White ● Other
nscore: *Numeric**	Participant's neuroticism score from the NEO-FFI-R test. Neuroticism scores range from 10-60 with lower scores representing low levels of neuroticism. Only the scores to the right of the colon were present among participants. The scores are defined by the following numeric values. ● -3.46436: 12 ● -3.15735: 13 ● -2.75696: 14 ● -2.52197: 15 ● -2.42317: 16 ● -2.34360: 17 ● -2.21844: 18 ● -2.05048: 19 ● -1.86962: 20 ● -1.69163: 21 ● -1.55078: 22 ● -1.43907: 23 ● -1.32828: 24 ● -1.19430: 25 ● -1.05308: 26 ● -0.92104: 27 ● -0.79151: 28 ● -0.67825: 29 ● -0.58016: 30 ● -0.46725: 31 ● -0.34799: 32 ● -0.24649: 33 ● -0.14882: 34 ● -0.05188: 35 ● 0.04257: 36 ● 0.13606: 37 ● 0.22393: 38 ● 0.31287: 39 ● 0.41667: 40

	• 0.52135: 41 • 0.62967: 42 • 0.73545: 43 • 0.82562: 44 • 0.91093: 45 • 1.02119: 46 • 1.13281: 47 • 1.23461: 48 • 1.37297: 49 • 1.49158: 50 • 1.60383: 51 • 1.72012: 52 • 1.83990: 53 • 1.98437: 54 • 2.12700: 55 • 2.28554: 56 • 2.46262: 57 • 2.61139: 58 • 2.82196: 59 • 3.27393: 60
escore: *Numeric**	Participant's extraversion score from the NEO-FFI-R test. Extraversion scores range from 10-60 with lower scores representing introversion. Only the scores to the right of the colon were present among participants. The scores are defined by the following numeric values. • -3.27393: 16 • -3.00537: 18 • -2.72827: 19 • -2.53830: 20 • -2.44904: 21 • -2.32338: 22 • -2.21069: 23 • -2.11437: 24 • -2.03972: 25 • -1.92173: 26 • -1.76250: 27 • -1.63340: 28 • -1.50796: 29 • -1.37639: 30 • -1.23177: 31 • -1.09207: 32 • -0.94779: 33 • -0.80615: 34

	• -0.69509: 35 • -0.57545: 36 • -0.43999: 37 • -0.30033: 38 • -0.15487: 39 • 0.00332: 40 • 0.16767: 41 • 0.32197: 42 • 0.47617: 43 • 0.63779: 44 • 0.80523: 45 • 0.96248: 46 • 1.11406: 47 • 1.28610: 48 • 1.45421: 49 • 1.58487: 50 • 1.74091: 51 • 1.93886: 52 • 2.12700: 53 • 2.32338: 54 • 2.57309: 55 • 2.85950: 56 • 3.00537: 58 • 3.27393: 59
oscore: *Numeric**	Participant's openness to experience score from the NEO-FFI-R test. Openness to experience scores range from 10-60 with lower scores representing low level of openness to experience. Only the scores to the right of the colon were present among participants. The scores are defined by the following numeric values. • -3.27393: 24 • -2.85950: 26 • -2.63199: 28 • -2.39883: 29 • -2.21069: 30 • -2.09015: 31 • -1.97495: 32 • -1.82919: 33 • -1.68062: 34 • -1.55521: 35 • -1.42424: 36 • -1.27553: 37 • -1.11902: 38

	• -0.97631: 39 • -0.84732: 40 • -0.71727: 41 • -0.58331: 42 • -0.45174: 43 • -0.31776: 44 • -0.17779: 45 • -0.01928: 46 • 0.14143: 47 • 0.29338: 48 • 0.44585: 49 • 0.58331: 50 • 0.72330: 51 • 0.88309: 52 • 1.06238: 53 • 1.24033: 54 • 1.43533: 55 • 1.65653: 56 • 1.88511: 57 • 2.15324: 58 • 2.44904: 59 • 2.90161: 60
ascore: *Numeric**	Participant's agreeableness score from the NEO-FFI-R test. Agreeable scores range from 10-60 with lower scores representing lower levels of agreeableness. Only the scores to the right of the colon were present among participants. The scores are defined by the following numeric values. • -3.46436: 12 • -3.15735: 16 • -3.00537: 18 • -2.90161: 23 • -2.78793: 24 • -2.70172: 25 • -2.53830: 26 • -2.35413: 27 • -2.21844: 28 • -2.07848: 29 • -1.92595: 30 • -1.77200: 31 • -1.62090: 32 • -1.47955: 33 • -1.34289: 34 • -1.21213: 35

	• -1.07533: 36 • -0.91699: 37 • -0.76096: 38 • -0.60633: 39 • -0.45321: 40 • -0.30172: 41 • -0.15487: 42 • -0.01729: 43 • 0.13136: 44 • 0.28783: 45 • 0.43852: 46 • 0.59042: 47 • 0.76096: 48 • 0.94156: 49 • 1.11406: 50 • 1.2861: 51 • 1.45039: 52 • 1.61108: 53 • 1.81866: 54 • 2.03972: 55 • 2.23427: 56 • 2.46262: 57 • 2.75696: 58 • 3.15735: 59 • 3.46436: 60
cscore: *Numeric**	Participant's conscientiousness score from the NEO-FFI-R test. Conscientiousness scores range from 10-60 with lower scores representing lower levels of conscientiousness. Only the scores to the right of the colon were present among participants. The scores are defined by the following numeric values. • -3.46436: 17 • -3.15735: 19 • -2.90161: 20 • -2.72827: 21 • -2.57309: 22 • -2.42317: 23 • -2.30408: 24 • -2.18109: 25 • -2.04506: 26 • -1.92173: 27 • -1.78169: 28 • -1.64101: 29

	• -1.51840: 30 • -1.38502: 31 • -1.25773: 32 • -1.13788: 33 • -1.01450: 34 • -0.89891: 35 • -0.78155: 36 • -0.65253: 37 • -0.52745: 38 • -0.40581: 39 • -0.27607: 40 • -0.14277: 41 • -0.00665: 42 • 0.12331: 43 • 0.25953: 44 • 0.41594: 45 • 0.58489: 46 • 0.7583: 47 • 0.93949: 48 • 1.13407: 49 • 1.30612: 50 • 1.46191: 51 • 1.63088: 52 • 1.81175: 53 • 2.04506: 54 • 2.33337: 55 • 2.63199: 56 • 3.00537: 57 • 3.46436: 59
impulsive: *Numeric**	Participant's impulsiveness score from the BIS-11 test. Impulsiveness scores range from 30-120 with lower scores representing lower levels of impulsiveness. The dataset only provided each participant's impulsiveness score as the number of standard deviations away from the mean impulsiveness score. • The number of standard deviations from the mean score range from -2.55524 - 2.90161.
SS: *Numeric**	Participant's sensation-seeking score from the ImpSS test. Sensation-seeking scores range from 0-19 with lower scores representing lower levels of sensation-seeking. The dataset only provided each participant's sensation

	seeking score as the number of standard deviations away from the mean sensation-seeking score. ● The number of standard deviations from the mean score range from -2.07848 - 1.92173
alcohol: class *Nominal*†	Participant's alcohol consumption frequency. Its values fall into the following classes: ● CL0: Never used ● CL1: Used over a decade ago ● CL2: Used in the last decade ● CL3: Used in the last year ● CL4: Used in the last month ● CL5: Used in the last week ● CL6: Used in the last day
amphet: class *Nominal*†	Participant's amphetamine consumption frequency. Its values fall into the following classes: ● CL0: Never used ● CL1: Used over a decade ago ● CL2: Used in the last decade ● CL3: Used in the last year ● CL4: Used in the last month ● CL5: Used in the last week ● CL6: Used in the last day
amyl: class *Nominal*†	Participant's amyl nitrite consumption frequency. Its values fall into the following classes: ● CL0: Never used ● CL1: Used over a decade ago ● CL2: Used in the last decade ● CL3: Used in the last year ● CL4: Used in the last month ● CL5: Used in the last week ● CL6: Used in the last day
benzos: class *Nominal*†	Participant's benzodiazepine consumption frequency. Its values fall into the following classes: ● CL0: Never used ● CL1: Used over a decade ago ● CL2: Used in the last decade ● CL3: Used in the last year

	• CL4: Used in the last month • CL5: Used in the last week • CL6: Used in the last day
caff: class *Nominal*†	Participant's caffeine consumption frequency. Its values fall into the following classes: • CL0: Never used • CL1: Used over a decade ago • CL2: Used in the last decade • CL3: Used in the last year • CL4: Used in the last month • CL5: Used in the last week • CL6: Used in the last day
cannabis: class *Nominal*†	Participant's cannabis consumption frequency. Its values fall into the following classes: • CL0: Never used • CL1: Used over a decade ago • CL2: Used in the last decade • CL3: Used in the last year • CL4: Used in the last month • CL5: Used in the last week • CL6: Used in the last day
choc: class *Nominal*†	Participant's chocolate consumption frequency. Its values fall into the following classes: • CL0: Never used • CL1: Used over a decade ago • CL2: Used in the last decade • CL3: Used in the last year • CL4: Used in the last month • CL5: Used in the last week • CL6: Used in the last day
coke: class *Nominal*†	Participant's cocaine consumption frequency. Its values fall into the following classes: • CL0: Never used • CL1: Used over a decade ago • CL2: Used in the last decade • CL3: Used in the last year • CL4: Used in the last month • CL5: Used in the last week • CL6: Used in the last day
crack: class *Nominal*†	Participant's crack consumption frequency. Its

	values fall into the following classes: ● CL0: Never used ● CL1: Used over a decade ago ● CL2: Used in the last decade ● CL3: Used in the last year ● CL4: Used in the last month ● CL5: Used in the last week ● CL6: Used in the last day
ecstasy: class *Nominal*†	Participant's ecstasy consumption frequency. Its values fall into the following classes: ● CL0: Never used ● CL1: Used over a decade ago ● CL2: Used in the last decade ● CL3: Used in the last year ● CL4: Used in the last month ● CL5: Used in the last week ● CL6: Used in the last day
heroin: class *Nominal*†	Participant's heroin consumption frequency. Its values fall into the following classes: ● CL0: Never used ● CL1: Used over a decade ago ● CL2: Used in the last decade ● CL3: Used in the last year ● CL4: Used in the last month ● CL5: Used in the last week ● CL6: Used in the last day
ketamine: class *Nominal*†	Participant's ketamine consumption frequency. Its values fall into the following classes: ● CL0: Never used ● CL1: Used over a decade ago ● CL2: Used in the last decade ● CL3: Used in the last year ● CL4: Used in the last month ● CL5: Used in the last week ● CL6: Used in the last day
legalh: class *Nominal*†	Participant's legal highs consumption frequency. Its values fall into the following classes: ● CL0: Never used ● CL1: Used over a decade ago ● CL2: Used in the last decade

	• CL3: Used in the last year • CL4: Used in the last month • CL5: Used in the last week • CL6: Used in the last day
lsd: class *Nominal*†	Participant's LSD consumption frequency. Its values fall into the following classes: • CL0: Never used • CL1: Used over a decade ago • CL2: Used in the last decade • CL3: Used in the last year • CL4: Used in the last month • CL5: Used in the last week • CL6: Used in the last day
meth: class *Nominal*†	Participant's methadone consumption frequency. Its values fall into the following classes: • CL0: Never used • CL1: Used over a decade ago • CL2: Used in the last decade • CL3: Used in the last year • CL4: Used in the last month • CL5: Used in the last week • CL6: Used in the last day
mushrooms: class *Nominal*†	Participant's magic mushrooms consumption frequency. Its values fall into the following classes: • CL0: Never used • CL1: Used over a decade ago • CL2: Used in the last decade • CL3: Used in the last year • CL4: Used in the last month • CL5: Used in the last week • CL6: Used in the last day
nicotine: class *Nominal*†	Participant's nicotine consumption frequency. Its values fall into the following classes: • CL0: Never used • CL1: Used over a decade ago • CL2: Used in the last decade • CL3: Used in the last year • CL4: Used in the last month • CL5: Used in the last week • CL6: Used in the last day

semer: class *Nominal*†	Participant's semeron consumption frequency. Semeron is a made-up drug that was placed in the dataset to detect participants that lied about their drug consumption. Its values fall into the following classes: ● CL0: Never used ● CL1: Used over a decade ago ● CL2: Used in the last decade ● CL3: Used in the last year ● CL4: Used in the last month ● CL5: Used in the last week ● CL6: Used in the last day
vsa: class *Nominal*†	Participant's volatile substance abuse consumption frequency. Its values fall into the following classes: ● CL0: Never used ● CL1: Used over a decade ago ● CL2: Used in the last decade ● CL3: Used in the last year ● CL4: Used in the last month ● CL5: Used in the last week ● CL6: Used in the last day

*Converted to nominal for analyzing the Frequency dataset; †Removed from both transformed datasets; †Removed for analyzing Legal/Illegal dataset.

Table 1. Original Cleaned Dataset Attributes

Data preparation

We began the data preparation process by cleaning our original dataset. First, we converted the dataset from its .data format into an .arff file format for Weka accessibility. We completed this process by using Visual Studio Code to create the file title, its list of attributes, and the values for each instance in the proper syntax. Next, we went through the dataset and deleted all instances that claimed using the fictitious drug, semeron, at any point in time. We wanted to remove these instances in order to eliminate any possible noise in the form of inaccurate information in the dataset. Lastly, we changed our demographic attributes (age, gender, highest level of education, country of current residence, and ethnicity) from numeric to nominal attributes for readability and explainability purposes. We accomplished this with R using a nested ifelse() function to convert each numeric value to its respective nominal attribute value (indicated on the dataset's web page), and classified all instances from the indicated column as a factor using the as.factor() function. By making the values into factors, the attribute becomes a functional nominal attribute in Weka. The R code is displayed below in Figure 1.

```
drug_data$age <- ifelse(drug_data$age == -0.95197, "18-24",
                        ifelse(drug_data$age == -0.07854, "25-34",
                               ifelse(drug_data$age == 0.49788, "35-44",
                                      ifelse(drug_data$age == 1.09449, "45-54",
                                             ifelse(drug_data$age == 1.82213, "55-64",
                                                    ifelse(drug_data$age == 2.59171, "65-", drug_data$alcohol))))))

drug_data$age <- as.factor(drug_data$age)
```

Figure 1. R code for converting the numeric age attribute into a nominal attribute expressed as age ranges in our original cleaned dataset.

We then performed two forms of data transformation to simplify the dataset and improve the class attribute prediction accuracy. The original dataset includes eighteen different class attributes explaining the drug usage frequency for legal and illegal substances. The first method sought to combine the drug usage frequency attributes into a single drug classification class attribute. Originally we wanted the new class attribute to include the values "Used_Legal," "Used_Illegal," "Used_Both," and "Never_Used," but we found that the "Used_Illegal" and "Never_Used" classes covered less than ten instances in the entire dataset. To avoid overfitting, we deleted the singular instance that claimed to have never used any of the eighteen drugs and decided to make our drug usage class attribute contain only the values "Used_Legal," defining a participant that has only used legal substances at some point in his/her life, and "Used_Illegal," defining a participant that has used at least one illegal substance at some point in his/her life. This new dataset is referred to as the Legal/Illegal dataset. Using R, we identified the 12 illegal drug columns and saved a copy of them as a vector titled original_illegal_drug_cols. Next we used the lapply() and as.numeric functions to convert all of the illegal drug usage attribute values to numeric values ranging from 1-7 with 1 = CL0, 2 = CL1, 3 = CL2, 4 = CL3, 5 = CL4, 6 = CL5, and 7 = CL6. After the nominal to numeric conversion, we summed up all of the illegal drug usage values per participant and placed the sum in a new attribute column titled illegal_drug_use. If a participant's illegal_drug_use = 12, the participant never used an illegal drug because all the twelve illegal drug usage attributes had a value of 1. Any sum greater than 12 suggests that the participant used at least one illegal drug. Using this information, we created a new column called illegal_drug_classification and identified this attribute's values of "Illegal" and "Legal" for each instance using the ifelse() function. The function states that if the instance's sum is greater than 12 the illegal_drug_classification for the participant is "Illegal." Otherwise, illegal_drug_classification = "Legal." These rules are also displayed below in Figure 2.

```r
# Define which drugs are considered illegal
illegal_drugs <- c("amphet", "amyl", "benzos", "cannabis", "coke", "crack", "ecstasy", "heroin", "ketamine", "lsd", "meth", "mushrooms")

# Save a copy of the original illegal drug columns
original_illegal_drug_cols <- drug_data[, illegal_drugs]

# Convert selected columns to numeric
drug_data[, illegal_drugs] <- lapply(drug_data[, illegal_drugs], as.numeric)

# Calculate total usage of illegal drugs for each instance
drug_data$illegal_drug_usage <- rowSums(drug_data[, illegal_drugs], na.rm = TRUE)

# Create a new column to classify instances based on illegal drug use
drug_data$illegal_drug_classification <- ifelse(drug_data$illegal_drug_usage > 12, "Illegal", "Legal")

# Revert illegal drug columns back to their original form
drug_data[, illegal_drugs] <- original_illegal_drug_cols

# Convert "legal, illegal" column to factor
drug_data$illegal_drug_classification <- as.factor(drug_data$illegal_drug_classification)
```

Figure 2. R code for adding drug classification attribute column in the Legal/Illegal dataset.

Attribute	**Description**
illegal_drug_use: *Numeric**	The sum of the illegal drug usage values for each instance. ● Sums can range from 12-84
illegal_drug_classification: class *Nominal*	Classifies each participant based on whether he/she has only used legal drugs or used at least one illegal drug. Its values fall into the following classes: ● Legal ● Illegal

*Removed in Weka for analyzing the Legal/Illegal dataset

Table 2. Additional attributes in the Legal/Illegal dataset.

Our second data transformation changed the drug classification file created from our first pre-processing method. First, we used Weka to eliminate all instances with the “Legal” class value for the illegal_drug_classification attribute. We then plugged the file containing only instances that used at least one illegal drug into R and added class attribute columns describing each instance’s usage frequency for particular illegal drugs. We decided to create drug usage frequency attributes for the drugs cannabis, cocaine, crack, heroin, and methadone. Each substance’s drug usage frequency attributes included the class values “Never_Used,” “Over_decade,” “Last_decade,” “Last_year,” “Last_month,” “Last_week,” and “Last_day” correlating with the original drug usage frequency values CL0, CL1, …, CL6. This new dataset is referred to as the Frequency dataset. We added the new frequency columns and their class values again using the ifelse() and as.factor() functions. We see the displayed R code for this procedure below in Figure 3.

```r
drug_data$frequency_coke <- ifelse(drug_data$coke == "CL0", "Never_Used",
                           ifelse(drug_data$coke == "CL1", "Over_decade",
                                  ifelse(drug_data$coke == "CL2", "Last_decade",
                                         ifelse(drug_data$coke == "CL3", "Last_year",
                                                ifelse(drug_data$coke == "CL4", "Last_month",
                                                       ifelse(drug_data$coke == "CL5", "Last_week",
                                                              ifelse(drug_data$coke == "CL6", "Last_day", drug_data$coke)))))))

#Convert coke column to factor
drug_data$frequency_coke <- as.factor(drug_data$frequency_coke)
```

Figure 3. R code for adding the cocaine usage frequency column in the Frequency dataset.

Attributes	**Description**
frequency_cannabis: class *Nominal*	Participant's cannabis consumption frequency. Its values fall into the following classes: ● Never_Used ● Over_decade ● Last_decade ● Last_year ● Last_month ● Last_wek ● Last_day
frequency_coke: class *Nominal**	Participant's cocaine consumption frequency. Its values fall into the following classes: ● Never_Used ● Over_decade ● Last_decade ● Last_year ● Last_month ● Last_wek ● Last_day
frequency_crack: class *Nominal**	Participant's crack consumption frequency. Its values fall into the following classes: ● Never_Used ● Over_decade ● Last_decade ● Last_year ● Last_month ● Last_wek ● Last_day
frequency_heroin: class *Nominal**	Participant's heroin consumption frequency. Its values fall into the following classes: ● Never_Used ● Over_decade ● Last_decade

	• Last_year • Last_month • Last_wek • Last_day
frequency_meth: class *Nominal**	Participant's methadone consumption frequency. Its values fall into the following classes: • Never_Used • Over_decade • Last_decade • Last_year • Last_month • Last_wek • Last_day

*Removed in Weka for analyzing frequency_cannabis in the Frequency dataset.

Table 3. Additional attributes in the Frequency dataset.

After creating the two new datasets, we opened each in Weka and applied various filters to them before performing our data analysis. The Legal/Illegal dataset was heavily unbalanced in favor of illegal drug usage. We used Weka to create additional instances where legal drug usage was the class value. We additionally decided to remove certain attributes from the datasets before performing analysis. For both datasets, the attributes 'individual_id' and 'ethnicity' were removed. To avoid overfitting, we removed 'individual_id' because we did not want singular ID numbers used for making rules. We also removed 'ethnicity' because the dataset was heavily unbalanced in favor of 'White' and therefore would not contribute to making rules that are representative of all demographic backgrounds. From the Legal/Illegal dataset, we wanted to demonstrate relationships between personality traits and legal/illegal drug use. Therefore, all attributes except 'nscore', 'escore', 'oscore', 'ascore', 'cscore', 'impulsive', 'SS', and 'illegal_drug_classification' were removed. From the Frequency dataset, we chose to focus on the usage frequency of cannabis and determine if using cannabis more frequently could serve as a gateway to using other hard drugs. Therefore, we set 'frequency_cannabis' as the class attribute, and removed all drug use attributes except for those classified as hard drugs such as crack, cocaine, heroin, and methadone.

Once we applied the necessary filters and removed unnecessary attributes, we used Weka to split each of our Legal/Illegal and Frequency datasets into two separate datasets, one for training and creating the model and one for testing the accuracy of the model. The training sets contained 90% of our Legal/Illegal and Frequency datasets, and the testing sets contained the remaining 10%.

Data Analysis

Once our data was prepared, we had to decide which learning methods would be the most applicable for our purposes.

Cost-Sensitive Classifier

Due to the unbalanced nature of our dataset, it was clear that a cost-sensitive classifier would create the most accurate model. This classifier works by taking a base classifier, in our case it was JRip, then makes it 'cost more' to incorrectly predict an instance. JRip is a classifier that predicts the class attribute by making rules with a dataset's attributes. The algorithm creates a rule by iterating over each pair of attributes and their associated values. After each iteration, it determines whether the addition of the attribute value pair to the rule improves the classification accuracy when compared to the training dataset. After developing a complete rule, JRip then goes through the rule and removes any preconditions that are not predictive or result in overfitting. These steps are repeated to create a set of rules until the algorithm has reached the maximum number of iterations or reached a maximum accuracy threshold. We then tested the model by supplying a test set, our Legal/Illegal test set, to compare to our models and determine the classification accuracy.

Vote

Vote is an ensemble learning algorithm. It works by creating models from a collection of learning algorithms and voting on the most commonly predicted class attribute. Using the Frequency dataset, we used Weka's *Vote* algorithm using five base classifiers: J48, DecisionTable, Prism, JRip, and OneR. A classification was made using the 'Majority Vote' rule. This means that each base classifier would make a classification and the class value that was predicted the most is the one the model would predict. We tested the model by supplying a test set, our Frequency test set, to compare to our models and determine the classification accuracy.

Results

While demographic attributes were included in the dataset, we decided to place our focus for data analysis on personality test scores as they served as more predictive attributes in the data mining process.

Legal/Illegal Dataset

For the cost-sensitive classifier, we used an explicit cost matrix for misclassifying someone who has used an illegal drug as a legal drug-only user. We chose to make it cost more to misidentify illegal as legal because it is a federal crime to use illegal drugs. As seen in Table 4, the confusion matrix shows the benefits of the cost matrix as there are few instances misclassified. The overall accuracy of the model was 92.9%. The rules created suggested a relationship between various personality trait scores and 'illegal_drug_classification'. 'Impulsive', 'SS', 'ascore', and 'nscore' appeared to be the top 4 most predictive attributes found by the model since they were used in so many of the rules.

Classified as →	**Illegal**	**Legal**
Illegal	**167**	8
Legal	14	**121**

Table 4. Confusion matrix from Legal/Illegal Cost-Sensitive model.

After analyzing these rules produced by the model, we decided to compare personality trait scores with illegal drug use across all instances within the dataset. Figure 6 demonstrates the correlation between illegal drug usage and the personality trait score. The colors on the heatmap represent the correlation between personality scores and drug usage. Darker shades indicate a higher correlation between the personality score and drug usage. A higher (or lower) numeric personality score does not necessarily mean a higher correlation. Rather the correlation heat map shows if there is a tendency for higher or lower personality scores related to drug use. Figure 6 visualizes our machine learning discoveries that individuals who have high 'nscores', low 'ascores', high 'impulsive' scores, and high 'sensation seeking' scores are more prone to illegal drug usage. This aligns with research, specifically relating to neuroticism. Highly neurotic individuals "may be impulsive, pessimistic, or struggle to cope with stress," and they may also "struggle to resist distractions or temptations" (Laporte 2019). Emotional instability (reflected by a high 'nscore') is a common feature of many mental illnesses. We see the mental illness/illegal drug usage correlation in many studies such as one conducted by the National Institute for Drug Abuse which states "that individuals with severe, mild, or even subclinical mental disorders may use drugs as a form of self-medication" (NIDA 2021).

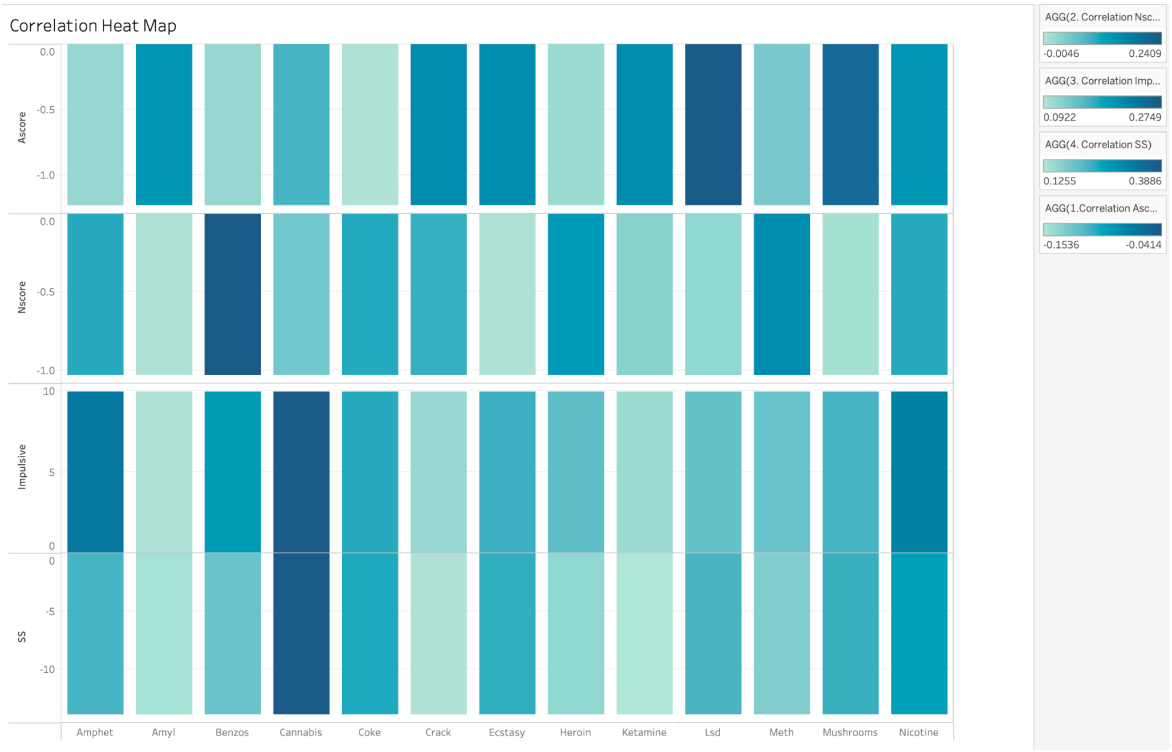

Figure 6. This heat map shows the correlation between personality scores and drug usage.

While addiction is a complex issue with many contributing factors, including genetics and environment, six personality traits can be best linked to an increased risk of addiction. Those six are: "impulsivity, nonconformity, anxiety, low tolerance for stress, sensation seeking, and blame shifting" (SJI). Using this model, we found that illegal usage is best predicted by the testable traits: impulsiveness, sensation-seeking, neuroticism, and agreeableness scores. Since the best attributes chosen for the model are loosely identical to the defined personality traits of an addict, our findings suggest that individuals who exhibit those personality traits and use illegal drugs are more susceptible to drug addiction.

Frequency Dataset

This dataset looked at the usage frequency of cannabis, crack, cocaine, heroin, and methadone. We specifically focused on predicting the frequency of using cannabis considering personality scores and the use of other drugs utilizing the *Vote* algorithm. Our *Vote* model had an accuracy of 49.73% overall. Our confusion matrix in Table 5 indicates that the majority of the cases in our dataset fall into the 'Last_Day' class. This led us to explore whether or not these instances with the greatest usage frequency also had a greater usage frequency for hard drugs.

Classified as →	**A**	**B**	**C**	**D**	**E**	**F**	**G**
A = Last_Day	**38**	0	0	0	0	20	0
B = Last_Decade	13	**3**	0	0	0	15	0
C= Last_Month	5	1	**2**	0	0	2	0
D= Last_Week	8	0	0	**4**	0	0	0
E = Last_Year	7	2	1	0	**1**	5	0
F= Never_Used	1	0	0	0	0	**44**	0
G= Over_Decade	3	0	0	0	0	11	**1**

Table 5. Confusion matrix for *Vote* model for Frequency dataset. Numbers highlighted indicate correctly classified instances.

The broad results of this model support the claim that cannabis is a ‘gateway’ drug. A gateway drug is best defined as a drug that “ may lead to the use of other types of high-risk addictive drugs” (SARC). One of our initial questions was related to identifying gateway drugs based on the frequencies of various drug usage. Figure 7 shows how the frequency usage of cannabis can be related to the frequency usage of other drugs including ‘coke’, ‘crack’, ‘heroin’, and ‘meth’. The most distinct trend is between the increased use of cannabis and cocaine. While the trends of crack, heroin, and methadone are not as sharp, there is still a noticeable increase with the increased use of cannabis.

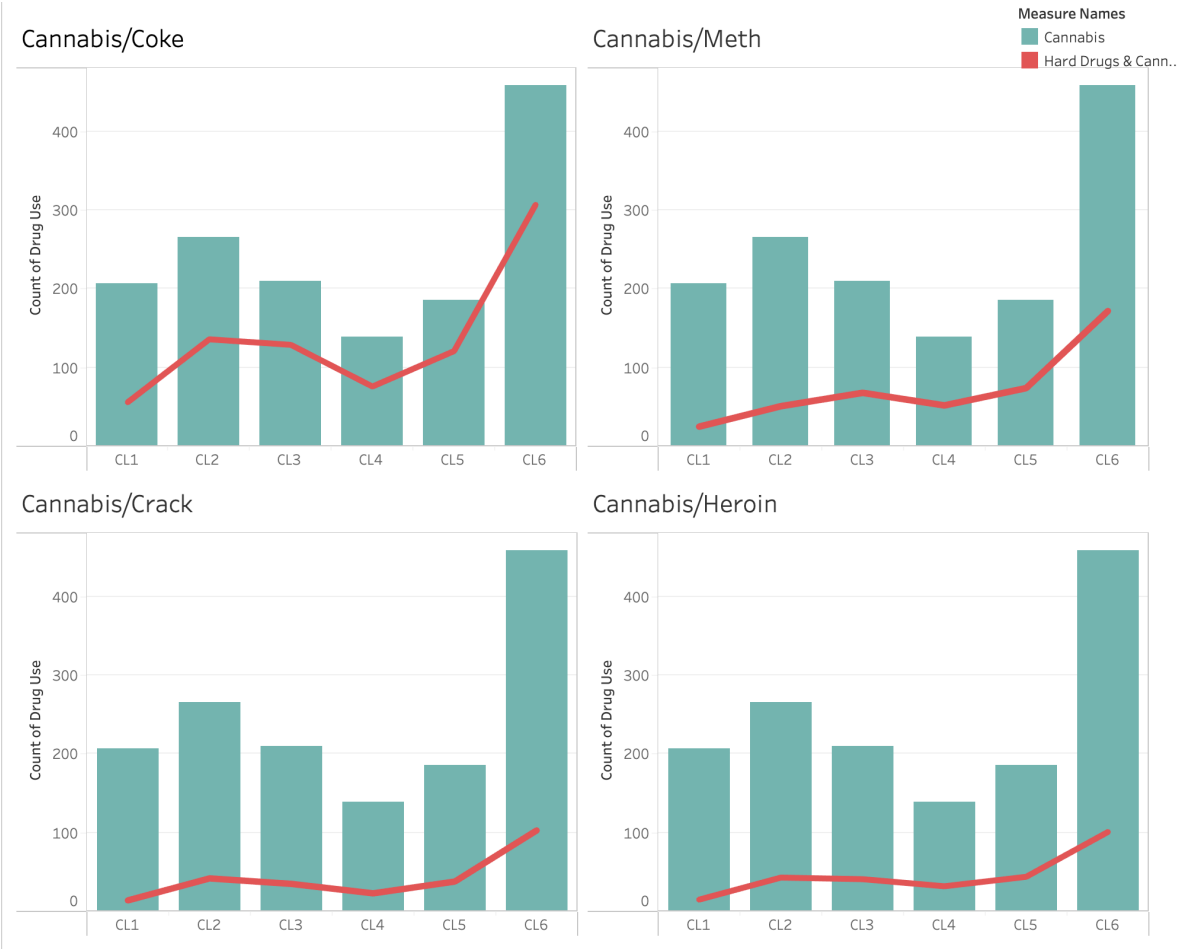

Figure 7. This visual depicts the overall frequency of cannabis consumption and contrasts it with the frequency of other drug use, among individuals who have also consumed cannabis.

Conclusion

Throughout this project, we have drawn connections between personality traits and drug usage. We have identified the levels of impulsiveness, agreeableness, neuroticism, and sensation-seeking personality traits aligned with mental illness to have the highest correlation with illegal hard drug usage. Knowing these correlations exist, we can use this dataset and machine learning to predict an individual's drug use habits to either enact preventative measures or create a proper addiction treatment plan. In addition, we have identified a 'gateway' drug, cannabis, by comparing its drug usage frequencies with hard drug usage frequencies across instances. With the increasing legalization of cannabis, these findings pose an argument about whether legalizing the drug will truly benefit society. The legalization of cannabis is intended to reduce illegal action around the substance, but its legalization could lead to the increased use of harder, illegal drugs, like cocaine. Years down the road, we may see that drugs like cocaine, methadone, and heroin are the new 'gateway' drugs that could worsen drug use and addiction. Therefore, using this dataset and machine learning tactics could help with identifying the most

influential gateway drugs and possibly proposing treatment plans or policy changes to improve the global drug addiction issue.

Data Source
Drug Consumption Quantified -
https://archive.ics.uci.edu/dataset/373/drug+consumption+quantified

Sources

University of Warwick. (n.d.). Personality and well-being results. Retrieved from https://warwick.ac.uk/services/dc/phdlife/wellbeing/potentialadvantage/personality_and_wellbeing_results/

NIDA. (2021, April 13). Why is there comorbidity between substance use disorders and mental illnesses? Retrieved from https://nida.nih.gov/publications/research-reports/common-comorbidities-substance-use-disorders/why-there-comorbidity-between-substance-use-disorders-mental-illnesses

San Antonio Recovery Center. (n.d.). What drugs are considered gateway drugs? Retrieved from https://www.sanantoniorecoverycenter.com/rehab-blog/what-drugs-are-considered-gateway-drugs/#:~:text=Gateway%20drugs%20are%20those%20that,using%20other%20drugs%20over%20time.

St. Joseph Institute. (n.d.). 6 personality traits linked to addiction. Retrieved from https://stjosephinstitute.com/6-personality-traits-linked-to-addiction/