Datamining for Business: Group Project

Section – BYGB7967-002

Presented on - 07/12/2021

Topic: Drug Consumption Risk Prediction
Project Report

GroupMembers

Asmita Tamhaney

Zahyra Ceballos

Arpita Choudhury

Sindoori Iyer

Abstract

The increasing rate of drug consumption has always been a serious problem. Data mining techniques present some advantages that can help us determine the probability of a respondent's drug consumption based on their personality traits. In this project, the drugs have been classified into different class labels used over different time frames.

There are 19 such drug classes and 12 attributes of respondents summarizing their demographic and behavioral aspects. We would be looking for a correlation, if present, in the behavioral and demographic information that would influence their usage pattern.

Lastly, we would be predicting if a respondent's consumption of one drug substance influences them to buy another drug.

Introduction

The increasing rate of unsupervised and illegal drug consumption has always been a serious problem. Drug addiction is a chronically relapsing disorder that has been characterized by the compulsive use of addictive substances despite adverse consequences to the individual and society. Addiction to drugs and alcohol is increasingly becoming a worldwide trend in lifestyle that is prevalent in rich and poor countries alike. Addiction to alcohol, drugs and cigarette smoking is now regarded as a major public health problem. Global drug consumption is not declining, even though its consequences are well known. In this sense, drinking and consuming hard substances endangers the lives of many. It can lead to physical and psychological problems, later addiction, and other drug-related problems that can last throughout their lives.

One of the most common users are teenagers who assume that consumption of psychoactive substances makes social situations more enjoyable. Cannabis, on the other hand, is widely consumed as a stress-coping strategy. People who suffer from social anxiety and depression might be more prone to drug use as it helps them relieve those symptoms. At times, drug consumption can solely be out of curiosity and social pressure. While drugs can be a form of entertainment, it does come with a lot of risks. The use of drugs by teenagers may have lasting brain changes and put the user at increased risk of dependence. Use and misuse of alcohol and banned drugs and prescription drugs cost Americans more than seven hundred billion dollars a year in health care costs, crime, and lost productivity. People of all ages suffer the damaging consequences of drug consumption and addiction. Adults who consume drugs may have trouble paying attention and thinking clearly. Adults who are parents and are inclined to drug consumption harm the wellbeing and development of children at home and may set the stage for drug consumption use in the next generation.

Every year, banned and prescription drug overdoses cause tens of thousands of deaths, alcohol contributes to the death of more than 90,000 Americans. America's Drug Overdose Epidemic is an ongoing issue that started in the late 90's and has increasingly gotten worse in recent years. According to the National Institute of Drug Abuse, in 2019, nearly 50,000 people in the United States died from opioid-involved overdoses for the misuse of drugs such as prescription pain relievers, heroin, and synthetic opioids (fentanyl).

In the period between March 2020 and March 2021 drug overdoses increased 29.6% as reported by the Centers for Disease Control and Prevention (CDC). And the COVID-19 pandemic has been another factor to worsen this public health crisis in the United States. The impact of the drug crisis has been so severe in American society that its effects have been felt socially, economically and even demographically, given that the opioid mortality rate contributed to a historic, three-year decline in life expectancy in the United States.

On the other hand, in the early 20th century, psychologist Carl Jung developed a theory of personality types to describe how different individuals function in the world. In the 1940s, Katherine Briggs and Isabel Briggs Myers expanded these ideas into a widely used personality test called the Myers-Briggs Tyler Indicator (MBTI). There were 4 major personality traits that are speculated to be determining factors of addiction. There is only limited evidence around personality types and substance abuse. There is no strong evidence, however, this method has been widely used in the field of psychology and proves as a strong indicator.

Over the years, significant amounts of data has been gathered about this phenomenon and performing advanced analysis on it, for prevention purposes that could strengthen public health response efforts, is highly relevant. The data collected could be used for predictive analysis, which allows creating individual risk profiles based on their personality traits and demographic information. This would facilitate a more focused and preventive response to the problem by state authorities.

Project Summary

In this assignment, with the help of data mining techniques we are trying to determine the probability of a respondent's drug consumption based on their personality traits. The data set chosen for this project consists of the drug classes used over different time frames. The drug classes are divided into 7 class labels, e.g. drug substance is used for a decade it means that it is consumed by the respondent for more than a decade. There are 19 such drug classes (cannabis, ecstasy, cocaine, etc) and 12 attributes of respondents summarizing their demographic and behavioral aspects. Our goal is to look for a correlation, if present, in the behavioral and demographic information that would influence their usage pattern. We would also be predicting if a respondent 's consumption of one drug substance influences them to buy another drug.

Data Description

The data was collected by surveying 1884 participants across the United States. Observations have been recorded based on 32 different variables.

The following are the Demographic attributes of each participant:

1. ID: Each participant is given an unique ID, to avoid duplication.

2. Age: Age of the participant. It has been categorized in ranges.

   Categories: 18-24, 25-34, 35-44, 45-54, 55-64, 65+

3. Gender: Gender of the Participant

   Categories: Male, Female

4. Education: Level of Education acquired by each participant.

   Categories: Left School before 16, Left School at 16, Left School at 17, Left School at 18, Some college or University no certificate or degree, University Degree, Professional certificate/ Diploma, Masters Degree, Doctorate Degree.

5. Country: Country of the participant

   Categories: USA, UK, New Zealand, Canada, Australia, Republic of Ireland, Other.

6. Ethnicity: Ethnicity of the participant

   Categories: Asian, Black, Mixed – Black/ Asian, Mixed – White/Asian, Mixed-White/ Black, Other Black

The next set of attributes are the personality traits of each participant: The Big Five personality traits is a suggested taxonomy, or grouping, for personality traits, developed from the 1980s onwards in psychological trait theory.

a. Nscore: Neuroticism is the tendency to experience negative emotions, such as anger, anxiety, or depression.

b. Escore: Extraversion is characterized by breadth of activities (as opposed to depth), surgency from external activity/situations, and energy creation from external means.

c. Oscore: Openness to experience is a general appreciation for art, emotion, adventure, unusual ideas, imagination, curiosity, and variety of experience. People who are open to experience are intellectually curious, open to emotion, sensitive to beauty and willing to try new things.

d. Ascore: The agreeableness trait reflects individual differences in general concern for social harmony. Agreeable individuals value getting along with others.

e. Cscore: Conscientiousness is a tendency to display self-discipline, act dutifully, and strive for achievement against measures or outside expectations. It is related to the way in which people control, regulate, and direct their impulses.

The last two personality traits are risk factors for hazardous and maladaptive behavior.

f. Impulsive: Impulsivity is a related yet distinct construct that reflects deficits in perseverance, planning, and inhibitory control. Impulsiveness measured by Barrent Impulsive Scale. It is a questionnaire designed to assess the personality/behavioral construct of impulsiveness. It is the most widely cited instrument for the assessment of impulsiveness and has been used to advance our understanding of this construct and its relationship to other clinical phenomena for 50 years. Measurement: Rarely/Never = 1, Occasionally = 2, Often = 3, Almost Always/Always = 4

g. SS: Sensation seeking is a personality trait that reflects the tendency to pursue and enjoy novel and stimulating experiences. Sensation Seeking measured by ImpSS. The ImpSS scale is a 19-question true-false scale assessing various personality characteristics and behaviors related to impulsivity and sensation seeking, and it is scored by summing the items that are consistent with impulsivity or sensation seeking.

Descriptive Statistics of Attributes:

Attributes	Datatypes	Descriptive Statistics			
		Min	Max	Mean	Standard Deviation
Nscore (Real) - NEO-FFR Neuroticism	Numeric	-3.464	3.274	0	0.998
Escore (Real) - NEO-FFR Extraversion	Numeric	-3.274	3.274	0	0.998
Oscore (Real) - NEO-FFR Openness to experience	Numeric	-3.274	2.902	0	0.996
Ascore (Real) - NEO-FFR Agreeableness	Numeric	-3.464	3.464	0	0.997
Cscore (Real) - NEO-FFR Conscientiousness	Numeric	-3.464	3.464	0	0.998
Impulsive (Real) measured by BIS-11	Numeric	-2.555	2.902	0.007	0.955
SS (Real) - sensation seeing measured by ImpSS	Numeric	-2.078	1.922	0.003	0.964

Drug Substances under study:

Finally, participants are questioned about their use of 18 different drugs and one fictitious drug named Semer to identify over claimers. The following are the 18 drugs: alcohol, amphetamines, amyl nitrite, benzodiazepine, caffeine, cannabis, chocolate, cocaine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse.

Drugs excluded from study - Caffeine, Chocolate, Semer - Class of Fictitious Drugs Semeron (i.e control)

Drug Classes: 7 Class Labels
They have been categorized with the class system of CL0-CL6: CL0 = "Never Used", CL1 = "Used over a decade ago", CL2 = "Used in last decade", CL3 = "Used in last year", CL4 = "Used in last month", CL5 = "Used in last week", CL6 = "Used in last day".

Problem Statement

In the United States, deaths due to drug overdose hit its highest records last year and continues to increase at a rapid speed. In order to alleviate such huge losses effective measures need to be taken. Therefore, in this project we aim to:

1. Identify which of the different drug substances can be bought together or if there is any association between these substances. Use of Apriori association for this Market basket analysis of different drug substances in the study.

2. . Evaluate the risk of being a drug consumer/user for each of the drug class (drug substance) based on demographic and behavioral data.
    a. Asses if the data can be used to predict the class label (different usage pattern) for each of the different drug classes (drug substance) separately.
    b. Asses if the problem can be converted into binary classification and check if each respondent is User or Non-user for the drug.

3. Identify if there is any correlation between
    a. Predictability of binary drugs consumptions using Demographic and Behavioral attributes separately
    b. Demographic attributes (e.g. age, education, ethnicity, etc.) and drug substance consumption
    c. Behavioral attributes (e.g. Escore, Ascore, Impulsive Score, etc.) and drug substance consumption
    d. Correlation within Behavioral attributes

Methodology

Stage 1: Data Collection

There are various repositories available online. We have finalized on the Drug consumption data set collected from UCI Machine learning repository.

Stage 2: Data Processing

- We need to remove the unwanted data. Here we ignore the ID in the data.
- Create Binary class for drug consumption classes (User and Non user) using existing class labels.

Stage 3: Use Data Mining and Algorithms

- Use of Apriori Association Mining for different drug substances consumed
- Use of Classification Models like Decision Tree and Bayes Net for 7 class labels for every drug substance.
- Similar classification to be used for Binary class labels for every drug substance
- Correlation of Demographic and Behavioral attributes - individually and together with every drug substance

Stage 4: Measure for Evaluation

- Data was subjected to 5 fold cross validation
- Measures used – Overall accuracy, Confusion matrix, and graphs for analysis

Results and Discussions:

1. **Association - Apriori**

Another component of this project's problem statement was to identify which of the different drug substances can be bought together or if there is any association between these substances. For this objective, we conducted a Market Basket Analysis (MBA) using the Apriori algorithm.

Market Basket Analysis (MBA) is a data mining technique that allows for identification of associations between attributes. In this case, the tool was used to identify patterns of consumption between the different drug classes and find strong association rules between drugs that have deadly and highly dangerous interactions.

With a support of 0.2, a confidence of 0.9 and 16 cycles performed, the best rules found amongst the drug substances were interactions between synthetic drugs and cannabis that led to alcohol consumption. Based on these rules, Cannabis wasn't found to lead to consumption of synthetic drugs, but to be taken in company of them, and that combination would then lead to alcohol consumption. Furthermore, a deadly combination between benzodiazepines and alcohol was detected with a confidence of 95%.

1. Coke Binary=1 417 ==> **Alcohol Binary**=1 412 <conf:(0.99)> lift:(1.06) lev:(0.01) [25] conv:(5.02)
2. Cannabis Binary=1 Ecstasy Binary=1 477 ==> **Alcohol Binary**=1 465 <conf:(0.97)> lift:(1.05) lev:(0.01) [22] conv:(2.65)
3. Ecstasy Binary=1 517 ==> **Alcohol Binary**=1 503 <conf:(0.97)> lift:(1.05) lev:(0.01) [23] conv:(2.49)
4. Cannabis Binary=1 Legalh Binary=1 521 ==> **Alcohol Binary**=1 505 <conf:(0.97)> lift:(1.04) lev:(0.01) [21] conv:(2.21)
5. Cannabis Binary=1 Mushrooms Binary=1 416 ==> **Alcohol Binary**=1 402 <conf:(0.97)> lift:(1.04) lev:(0.01) [16] conv:(2)
6. Benzos Binary=1 Cannabis Binary=1 425 ==> **Alcohol Binary**=1 410 <conf:(0.96)> lift:(1.04) lev:(0.01) [15] conv:(1.92)
13. **Benzos Binary=1 535 ==> Alcohol Binary**=1 507 <conf:(0.95)> lift:(1.02) lev:(0.01) [10] conv:(1.33)

To further pivot the project, with a support of 0.2, a confidence of 0.9 and 16 cycles performed again, we removed all drug classes that are regularly used and only considered synthetic or "hard" drugs. Here we found more dangerous drug combinations, however, they were still leading to alcohol consumption and the lift levels weren't increasing. This was also the first discovery of "speedballing" patterns of drug abuse, where alcohol was mainly being used as a downer.

1. Coke Binary=1 Ecstasy Binary=1 301 ==> Alcohol Binary=1 299 <conf:(0.99)> lift:(1.07) lev:(0.01) [19] conv:(7.24)

2. Amphet Binary=1 Coke Binary=1 Ecstasy Binary=1 194 ==> Alcohol Binary=1 192 <conf:(0.99)> lift:(1.07) lev:(0.01) [12] conv:(4.67)

3. Coke Binary=1 417 ==> Alcohol Binary=1 412 <conf:(0.99)> lift:(1.06) lev:(0.01) [25] conv:(5.02)

4. Benzos Binary=1 Coke Binary=1 247 ==> Alcohol Binary=1 244 <conf:(0.99)> lift:(1.06) lev:(0.01) [14] conv:(4.46)

5. Ecstasy Binary=1 LSD Binary=1 276 ==> Alcohol Binary=1 272 <conf:(0.99)> lift:(1.06) lev:(0.01) [15] conv:(3.98).

Speedballing is a pattern of drug consumption that involves the combination of drugs with opposite effects to manipulate the "highs" and the "lows" produced by the substances. Drug co-use is attributable to higher mortality rates (Goodwin, 2020). "Speedballing" practices were identified. Taking drugs that have opposing effects, can result in an accidental overdose because the effects of each diminish but the amount consumed remains the same.

Furthermore, in order to increase interestingness and ensure that the popularity of the drug wouldn't over influence the results.

1. Amphet Binary=1 Coke Binary=1 246 ==> Ecstasy Binary=1 194 <conf:(0.79)> lift:(2.87) lev:(0.07) [126] conv:(3.37)

2. Meth Binary=1 320 ==> Benzos Binary=1 239 <conf:(0.75)> lift:(2.63) lev:(0.08) [148] conv:(2.79)

3. LSD Binary=1 380 ==> Ecstasy Binary=1 276 <conf:(0.73)> lift:(2.65) lev:(0.09) [171] conv:(2.63)

4. Coke Binary=1 417 ==> Ecstasy Binary=1 301 <conf:(0.72)> lift:(2.63) lev:(0.1) [186] conv:(2.59)

In conclusion, this apriori association showed that Cannabis is usually consumed in the company of other drugs, the consumption of Cannabis and other drugs has a strong incidence in resorting to alcohol consumption, alcohol is commonly combined with cocaine, which is highly dangerous and considered a lethal combination, Synthetic drugs show reincidence in alcohol consumption, and they also show lethal combinations like benzodiazepines, cocaine and alcohol. Lastly, when alcohol was not considered, lift values much greater than 1.0 were found, by taking an extremely popular item out of the "basket" more relevant associations were made.

2. Classification Models

The second problem statement asked to evaluate the risk of being a drug consumer/user for each of the drug class (drug substance) based on demographic and behavioral data.
First question was to assess if the data can be used to predict the class label (different usage pattern) for each of the different drug classes (drug substance) separately.

To analyze the above problem statement, Bayes Net and Decision Tree were made into use. After several trials and errors, the following parameters were finalized.

- Bayes Net: 5-fold cross validation and Max no. of parents = 5
- Decision Tree: 5-fold cross validation, default parameters

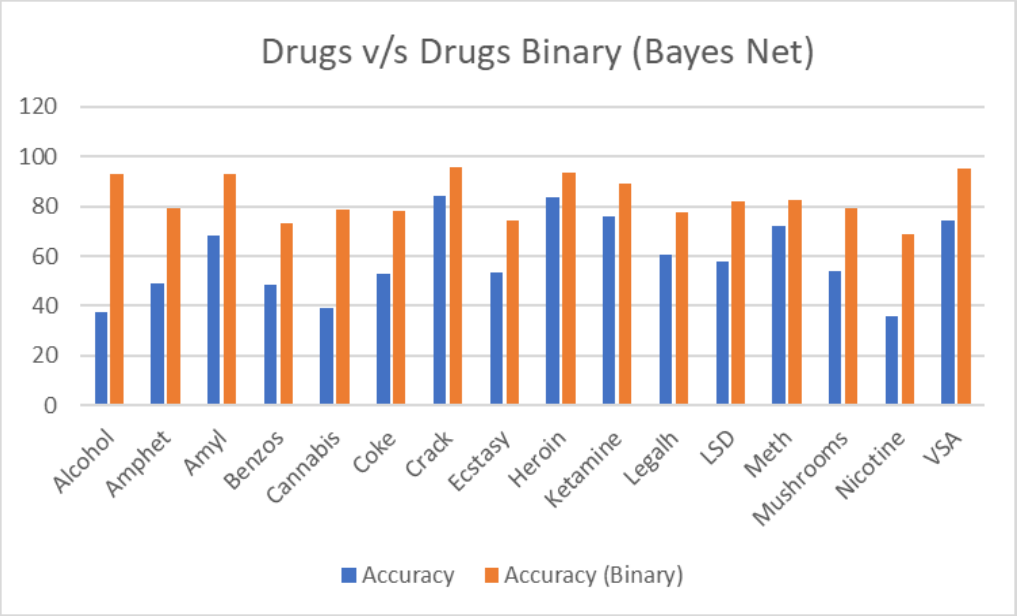
```
  a   b   c   d   e   f   g   <-- classified as
 18  55   2  64  13   0   4 |   a = CL4
 32 838   8 104  33   1   4 |   b = CL0
  0  90   0   7  15   0   1 |   c = CL1
 31  96   4 116  25   1   4 |   d = CL3
  9 131   6  48  36   1   3 |   e = CL2
  4   4   0  12   1   0   0 |   f = CL6
  5  16   0  37   5   0   0 |   g = CL5
```

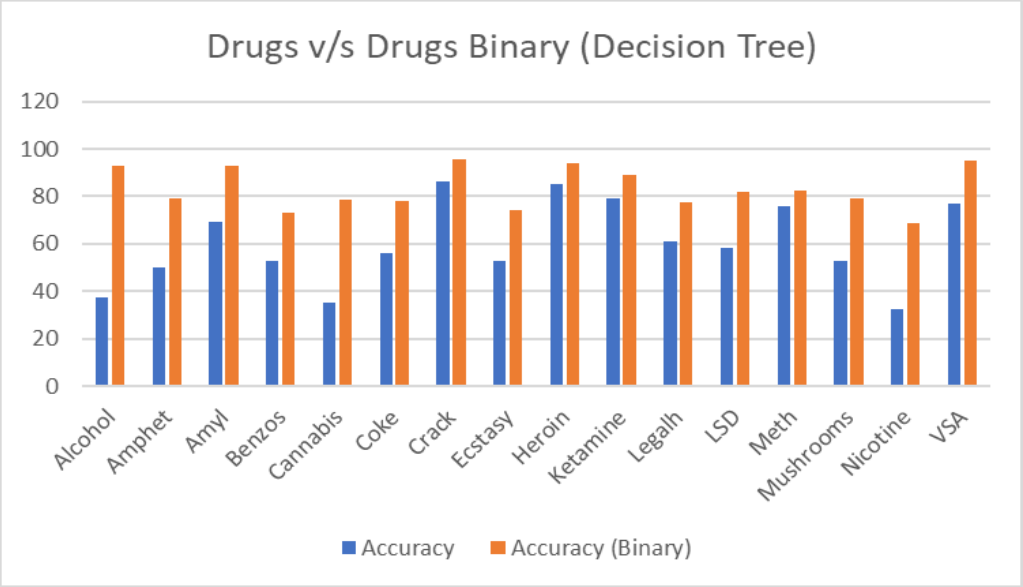
Though the model was able to predict the classes, the accuracy as a whole was very less. This was because of the reason that there was a confusion created between the nearby classes as seen in the above matrix.

This brings us to the second part of the problem statement which asks us to assess if the problem can be converted into binary classification and check if each respondent is User or Non-user for the drug. Keeping the parameters same as the above, the accuracy of the model was noted to increase significantly. For the below confusion matrix, only two class

labels were noted hence, the increase in accuracy. Both the confusion matrix taken as an example are for ecstasy and ecstasy binary.

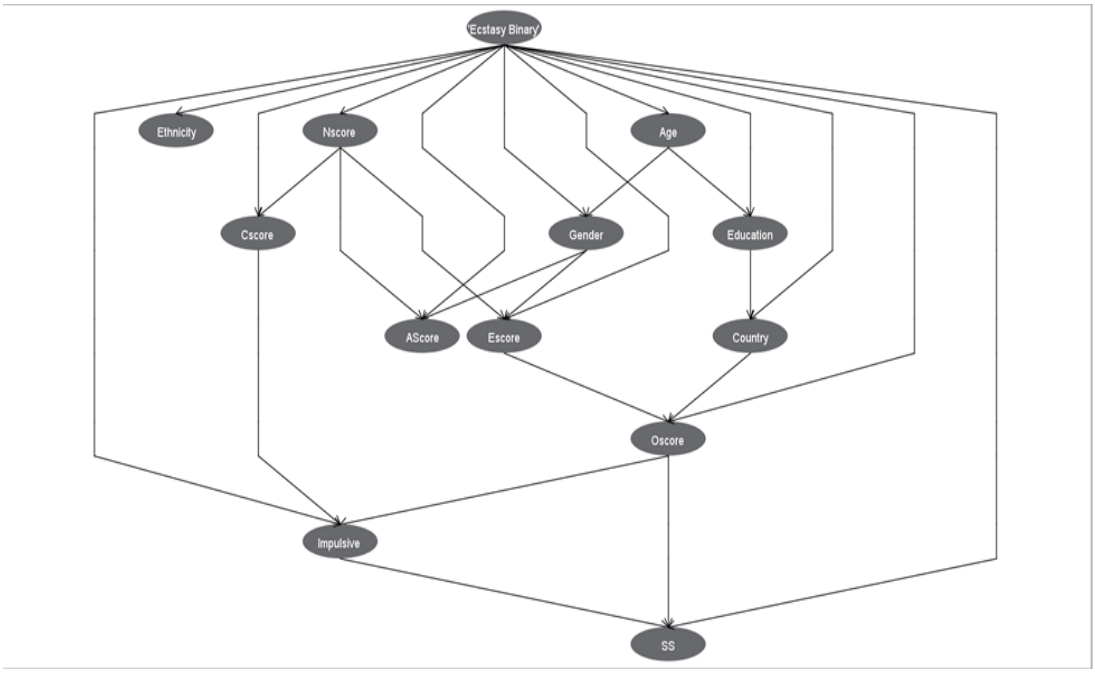
```
   a    b   <-- classified as
1137  230 |    a = 0
 217  300 |    b = 1
```





Out of the two classification models, Bayes Net was the preferred model for two reasons, first being for the visualization purpose as the decision tree model was an overfit. Secondly, Bayes net gives us the dependencies of attributes in predicting the consumption of a drug substance while decision tree provides us with the information gain on which feature could be more prominent for each drug substance consumption.

The below Bayes net tree helps in understanding the attributes contributing to the ecstasy consumption. For example: Age, education, country [demographic factors] along with Oscore and SS [behavioral attributes] lead to the consumption of that particular drug.



3. Correlation Analysis

Problem Statement:

To identify if there is any correlation between

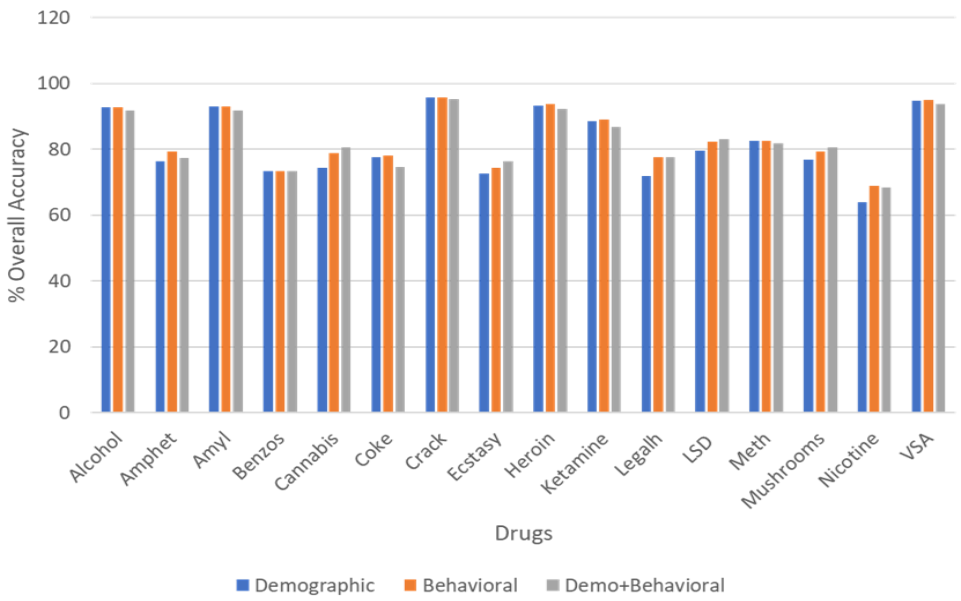
I. Predictability of binary drugs consumptions using Demographic and Behavioral attributes separately

After observing the Bayes Net graph for every binary drug substance it can be clearly seen that the consumption of these drug substances is highly dependent on the joint probability of Demographic and Behavioral attributes. This observation further led to evaluating the prediction of the drug substance consumption using demographic and behavioral attributes separately.

Classification algorithm Bayes Net was used to perform this evaluation and the parameter chosen were 5-fold cross validation and selecting the maximum number of parent nodes as five (5).

Graphical representation:

For overall accuracy of the model for each drug Vs Demographic and Behavioral attributes together and separately. (See Appendix 7.1 and 7.2)



Observations:

a. In general, classification performance is comparable in all three cases

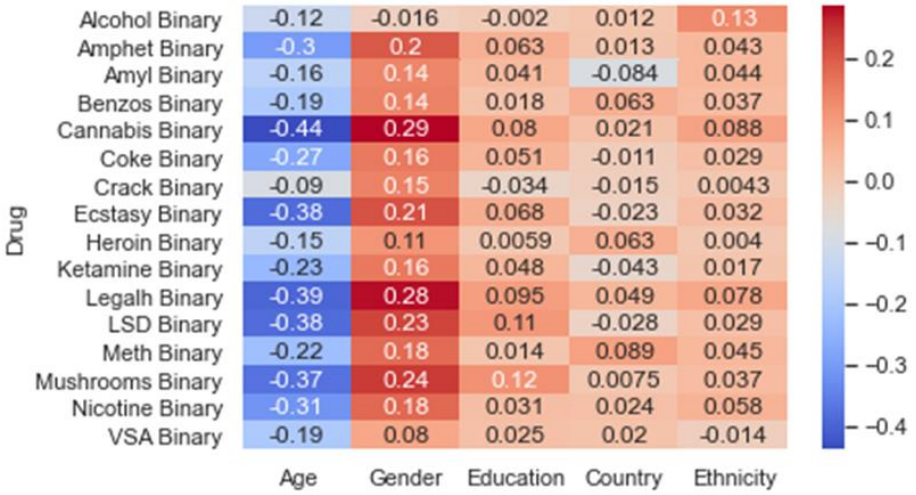
b. In case of Amphet, Cannabis, Ecstasy, Legalh, LSD, Mushrooms and Nicotine, behavioral features have performed better

These observations were the basis of our further evaluation of correlation analysis as mentioned below -

II. Demographic attributes (e.g. age, education, ethnicity, etc.) and drug substance consumption

Method Used: Corr (correlation) function in Pandas for codified demographic attributes and drug substances using Heatmap.

Since the demographic attributes are categorical variables, initial codification or conversion of these variables were performed. (See Appendix 7.3)



Heat map for Binary drug substances and demographic attributes

Observations**:**

a. Education, Country, Ethnicity show no significant correlation
b. Males are showing a higher correlation to drug consumption patterns. In order to verify this we performed cross tabulation for both categorical variables - each of the demographic attributes and binary drug classes. An example of this is explained below where in Cannabis (binary drug class - 0 = Non User and 1 = User) was evaluated against Gender using cross tabulation method in Pandas, Python. It can be clearly seen that Males were seen to have a higher inclination towards Cannabis consumption. A Similar trend was seen for drugs showing higher correlation in the heat map above.

Cannabis Binary	0	1
Gender		
F	0.61	0.39
M	0.33	0.67

Example of a drug substance - Cannabis and Gender, using cross-tabulation method

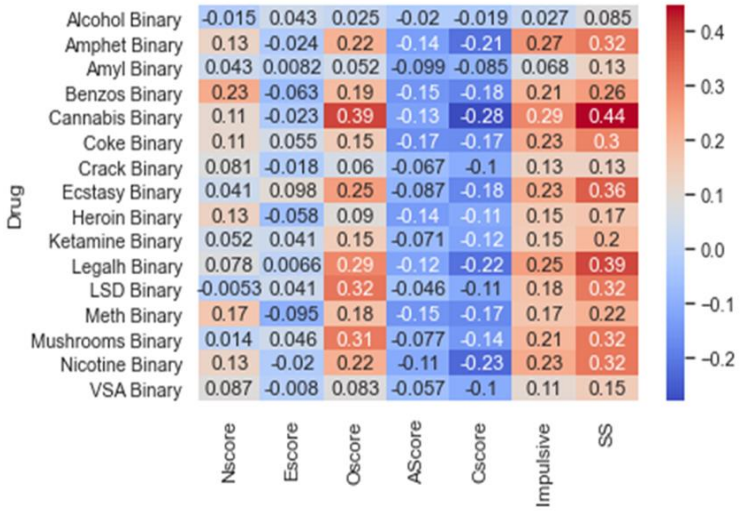
c. Lower age groups have a higher correlation to drug consumption. A cross tabulation method in Pandas, Python was performed on each of the demographic attributes and binary drug classes. An example of this is explained below where in Cannabis (binary drug class - 0 = Non User and 1 = User). It can be clearly seen that respondents in the age group of 18-24 years and 25-34 were seen to have a higher inclination towards Cannabis consumption. A Similar trend was seen for drugs showing higher correlation in the heat map above.

Cannabis Binary	0	1
Age		
18-24	0.17	0.83
25-34	0.51	0.49
35-44	0.65	0.35
45-54	0.74	0.26
55-64	0.72	0.28
65+	0.94	0.06
All	0.47	0.53

Example of a drug substance - Cannabis and Age, using cross-tabulation method

III. Behavioral attributes (e.g. Escore, Ascore, Impulsive Score, etc.) and drug substance consumption
Method used: Using Corr(correlation) function in Pandas using Heatmap

Drug	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS
Alcohol Binary	-0.01	4.31e-02	0.02	-0.02	-0.02	0.03	0.08
Amphet Binary	0.13	-2.39e-02	0.22	-0.14	-0.21	0.27	0.32
Amyl Binary	0.04	8.18e-03	0.05	-0.10	-0.08	0.07	0.13
Benzos Binary	0.23	-6.28e-02	0.19	-0.15	-0.18	0.21	0.26
Cannabis Binary	0.11	-2.26e-02	0.39	-0.13	-0.28	0.29	0.44



Heat map for Binary drug substances and behavioral attributes (scores)

Observations:

a. Cannabis is the only drug which has shown moderate correlation with behavioral attributes
    i. Higher Oscore (Openness to trying new things) and SS score (Sensation seeking) have shown to have positive correlation with Cannabis
    ii. Cscore (Conscientiousness – Self-discipline) is negatively correlated to Cannabis
b. Ecstasy and Legalh have also shown some positive correlation with SS score (Sensation seeking)
c. Nicotine has shown some negative correlation with Csore (Conscientiousness – Self-discipline)
d. Ascore (Agreeableness) and Escore (Extraversion) has shown relatively less correlation with the behavioral attributes

In general, it can be inferred that, Psychotropic drugs have some correlation with behavioral attributes (scores)

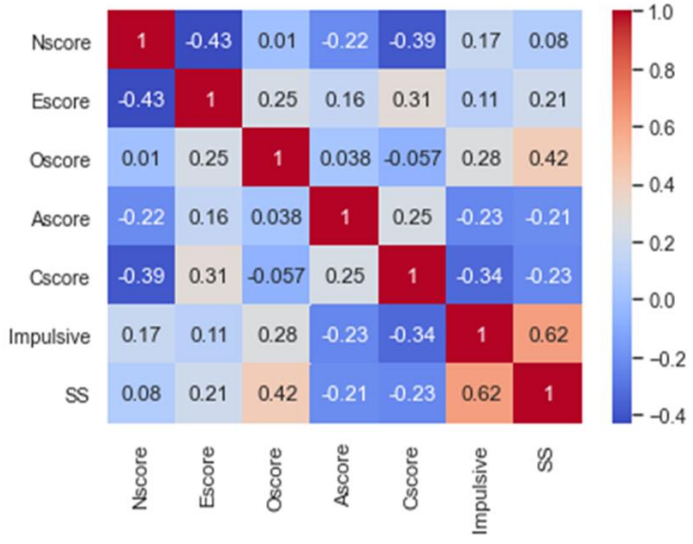
IV. Correlation within Behavioral attributes

In order to check the correlation pattern amongst the behavioral attributes (scores), we applied Pearson correlation method for the evaluation.

Method Used**:** Pearson correlation in Pandas with Corr function using heatmap

```
1 behavioural_corr = behaviour.corr()
2 behavioural_corr
```

	Nscore	Escore	Oscore	Ascore	Cscore	Impulsive	SS
Nscore	1.00	-0.43	0.01	-0.22	-0.39	0.17	0.08
Escore	-0.43	1.00	0.25	0.16	0.31	0.11	0.21
Oscore	0.01	0.25	1.00	0.04	-0.06	0.28	0.42
Ascore	-0.22	0.16	0.04	1.00	0.25	-0.23	-0.21
Cscore	-0.39	0.31	-0.06	0.25	1.00	-0.34	-0.23
Impulsive	0.17	0.11	0.28	-0.23	-0.34	1.00	0.62
SS	0.08	0.21	0.42	-0.21	-0.23	0.62	1.00



Heat map for Binary drug substances and behavioral attributes (scores)

Observations:

a. SS score (Sensation seeking) has shown a positive correlation with Impulsive score and Oscore (Openness)
b. Nscore (Neuroticism – e.g Anxitey, depression) has shown a negative correlation with Escore (Extraversion) and Cscore (Conscientiousness)
c. Impulsive score and Cscore (Conscientiousness) negatively correlated

In general, observed correlations validate known assumptions about behavioral traits

Conclusions:

1. Data mining techniques were proven useful at building several models that allowed them to predict behavior and classify instances of drug consumption.
2. Market Basket Analysis allowed us to discover hidden patterns within the data set, and allowed us to establish relationships between the attributes with high levels of confidence.
3. Binary classification of drugs showed better accuracy than drugs.
4. Bayes Net classification model helped in better visualization and identifying attributes dependencies for drug consumption.
5. Psychotropic drugs have some correlation with behavioral attributes
6. Drugs such as Cannabis were found to be part of almost every drug combination done by users, and it's also moderately correlated to the behavioral aspects of trying new things and sensation seeking. And negatively correlated with self-discipline
7. Contrary to what's usually believed, demographic data didn't draw strong enough correlations, beyond age group and gender (with younger people and males being slightly more likely to use)

Reference:

- E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," arXiv [Web Link], 2015
- Data source: https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29
- Un-quantified data from: https://www.kaggle.com/obeykhadija/drug-consumptions- uci/metadata
- Felter, C. (2021). The U.S. Opioid Epidemic. Retrieved from https://www.cfr.org/bacder/us-opioid-epidemic
- Understanding the Epidemic | CDC's Response to the Opioid Overdose Epidemic | CDC. (2021). Retrieved from https://www.cdc.gov/opioids/basics/epidemic.html
- https://www.simplypsychology.org/big-five-personality.html
- Goodwin, R., Moeller, S., Zhu, J., Yarden, J., Ganzhorn, S., & Williams, J. (2021). The potential role of cocaine and heroin co-use in the opioid epidemic in the United States. Addictive Behaviors, 113, 106680. doi: 10.1016/j.addbeh.2020.106680
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2678841/#:~:text=Finally%2C%20the%20ImpSS%20scale%20is,range%20from%200%20to%2019.

APPENDIX

1. **Abstract**
2. **Introduction**
3. **Data description**
4. **Methodology**
5. **Apriori Association**
6. **Classification (Drugs vs Drugs Binary)**

6.1 Drug and Demographic and Behavioral attributes together using Bayes Net

Target Feature	Method	Parameters	No. of Attributes	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS	Overall Accuracy (%)	Correctly classified instances	Incorrectly classified instances
Alcohol	Bayes Net	Cross validation - 5 folds Max Number of parents - 5	12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	37.42	705	1179
Amphet			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	48.89	921	963
Amyl			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	68.15	1284	600
Benzos			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	48.35	911	973
Cannabis			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	39.28	740	1144
Coke			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	52.76	994	890
Crack			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	84.39	1590	294
Ecstasy			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	53.5	1008	876
Heroin			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	83.6	1575	309
Ketamine			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	75.85	1429	455
Legalh			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	60.51	1140	744
LSD			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	57.86	1090	794
Meth			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	71.87	1354	530
Mushrooms			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	53.87	1015	869
Nicotine			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	35.77	674	1210
VSA			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	74.52	1404	480

6.2 Drug and Demographic and Behavioral attributes together using Decision Tree

Target Feature	Method	Parameters	No. of Attributes	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS	Overall Accuracy (%)	Correctly classified instances	Incorrectly classified instances
Alcohol	Decision Trees	Cross validation - 5 folds Default parameters	12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	37.26	702	1182
Amphet			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	50.27	947	937
Amyl			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	69.21	1304	580
Benzos			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	52.71	993	891
Cannabis			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	35.51	669	1215
Coke			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	56.21	1059	825
Crack			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	86.31	1626	258
Ecstasy			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	52.6	991	893
Heroin			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	85.14	1604	280
Ketamine			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	79.03	1489	395
Legalh			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	61.09	1151	733
LSD			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	58.55	1103	781
Meth			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	75.8	1428	456
Mushrooms			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	52.6	991	893
Nicotine			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	32.64	615	1269
VSA			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	77.18	1454	430

6.3 Binary Drug class with Demographic and Behavioral attributes together using Bayes Net

Target Feature	Method	Parameters	No. of Attributes	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS	Overall Accuracy (%)	Correctly classified instances	Incorrectly classified instances
Alcohol Binary	Bayes Net	Cross validation - 5 folds	12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	91.61	1726	158
Amphet Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	77.22	1455	429
Amyl Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	91.82	1730	154
Benzos Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	73.24	1380	504
Cannabis Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	80.63	1519	365
Coke Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	74.57	1405	479
Crack Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	95.11	1792	92
Ecstasy Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	76.27	1437	447
Heroin Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	92.09	1735	149
Ketamine Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	86.67	1633	251
Legalh Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	77.44	1459	425
LSD Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	83.06	1565	319
Meth Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	81.84	1542	342
Mushrooms Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	80.57	1518	366
Nicotine Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	68.36	1288	595
VSA Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	93.63	1764	120

6.4 Binary Drug class with Demographic and Behavioral attributes together using Decision Tree

Target Feature	Method	Parameters	No. of Attributes	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS	Overall Accuracy (%)	Correctly classified instances	Incorrectly classified instances
Alcohol Binary	Decision Trees	Cross validation - 5 folds	12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	92.78	1748	136
Amphet Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	79.29	1494	390
Amyl Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	92.94	1751	133
Benzos Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	73.4	1382	501
Cannabis Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	78.87	1486	398
Coke Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	78.02	1470	414
Crack Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	95.8	1805	79
Ecstasy Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	74.41	1402	482
Heroin Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	93.73	1766	118
Ketamine Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	88.95	1676	208
Legalh Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	77.44	1459	0.77
LSD Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	82.16	1548	336
Meth Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	82.43	1553	331
Mushrooms Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	79.19	1492	392
Nicotine Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	68.94	1299	585
VSA Binary			12	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	94.95	1789	95

7. Correlation

7.1 Binary Drug classes with Demographic attributes alone

Target Feature	Method	Parameters	No. of Attributes	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS	Overall Accuracy (%)	Correctly classified instances	Incorrectly classified instances
Alcohol Binary	Bayes Net	Cross validation - 5 folds Max Number of parents - 5	5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	92.52	1743	141
Amphet Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	78.29	1475	409
Amyl Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	92.73	1747	137
Benzos Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	74.47	1403	481
Cannabis Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	78.72	1483	401
Coke Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	76.27	1437	447
Crack Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	95.28	1795	89
Ecstasy Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	75.74	1427	457
Heroin Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	92.88	1750	134
Ketamine Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	88.32	1664	220
Legalh Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	78.76	1484	400
LSD Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	81.68	1539	345
Meth Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	81.52	1536	348
Mushrooms Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	81.58	1537	347
Nicotine Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	67.3	1268	616
VSA Binary			5	Y	Y	Y	Y	Y	N	N	N	N	N	N	N	94.69	1784	100

7.2 Binary Drug classes with Behavioral attributes alone

Target Feature	Method	Parameters	No. of Attributes	Age	Gender	Education	Country	Ethnicity	Nscore	Escore	Oscore	AScore	Cscore	Impulsive	SS	Overall Accuracy (%)	Correctly classified instances	Incorrectly classified instances
Alcohol Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	92.78	1748	136
Amphet Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	76.33	1438	446
Amyl Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	92.94	1751	133
Benzos Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	73.35	1382	502
Cannabis Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	74.42	1402	482
Coke Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	77.55	1461	423
Crack Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	95.81	1805	79
Ecstasy Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	72.56	1367	517
Heroin Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	93.15	1755	129
Ketamine Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	88.58	1669	215
Legalh Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	71.86	1354	530
LSD Binary			7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	79.51	1498	386
Meth Binary		Cross validation -	7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	82.64	1557	327
Mushrooms Binary		5 folds	7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	76.69	1445	439
Nicotine Binary		Max Number of	7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	63.85	1203	681
VSA Binary	Bayes Net	parents - 5	7	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	94.79	1786	98

7.3 Codification of each demographic (categorical) attributes for further simplification

```
#Codification of categorical varibales in Demograohic attributes
demographic["Age"] = pd.Categorical(demographic["Age"])
demographic["Age Code"] = demographic["Age"].cat.codes

demographic["Gender"] = pd.Categorical(demographic["Gender"])
demographic["Gender Code"] = demographic["Gender"].cat.codes

demographic["Education"] = pd.Categorical(demographic["Education"])
demographic["Education Code"] = demographic["Education"].cat.codes

demographic["Country"] = pd.Categorical(demographic["Country"])
demographic["Country Code"] = demographic["Country"].cat.codes

demographic["Ethnicity"] = pd.Categorical(demographic["Ethnicity"])
demographic["Ethnicity Code"] = demographic["Ethnicity"].cat.codes
```

```
1 demographic.to_csv("Codified Demo.csv")
2 demographic.head()
```

ountry	Ethnicity	Alcohol Binary	Amphet Binary	Amyl Binary	Benzos Binary	Caff Binary	...	Meth Binary	Mushrooms Binary	Nicotine Binary	Semer Binary	VSA Binary	Age Code	Gender Code	Education Code	Country Code	Ethnicity Code
UK	White	1	0	0	0	1	...	1	0	1	0	0	1	1	0	5	6
UK	White	1	0	0	0	1	...	0	0	0	0	0	2	1	6	5	6
UK	White	1	0	0	1	1	...	0	0	0	0	0	0	0	5	5	6
UK	White	1	0	0	0	1	...	0	0	0	0	0	2	0	0	5	6
anada	White	0	0	0	0	1	...	0	0	1	0	0	5	0	3	1	6

8. Results & Discussion

9. Conclusions