Drug Consumption Risk Prediction

Datamining for Business: Project Phase I – Project Proposal

Group Members:

Asmita Tamhaney

Zahyra Ceballos

Arpita Choudhury

Sindoori Iyer

FORDHAM UNIVERSITY

THE JESUIT UNIVERSITY OF NEW YORK

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. 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.

Keywords: data mining, influences, cannabis, ecstasy, cocaine

Introduction

- 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.
- Usage among teenagers and adults
- 90,000 deaths per year among Americans. Increase in America's Drug Overdose Epidemic since the 90's.
- Covid impact: Drug overdose increased by 29.6% between March 2020 and March 2021 reported by CDC.

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.

In this assignment, with the help of data mining techniques we are trying to present some advantages that can help us determine the probability of a respondent's drug consumption.

Data Description

Total of 1884 records/ participants.

Known Attributes			
1	ID	Numeric	Unique ID/ Numeric
2	Age	Numeric	Range/ 18-24, 25-34, 35-44, 45-54, 55-64, 65+
3	Gender	Nominal	M/F
4	Education	Nominal	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	Nominal	Countries/ USA, UK, New Zealand, Canada, Australia, Republic of Ireland, Other.
6	Ethnicity	Nominal	Categories/ Asian, Black, Mixed – Black/ Asian, Mixed – White/Asian, Mixed-White/ Black, Other Black.

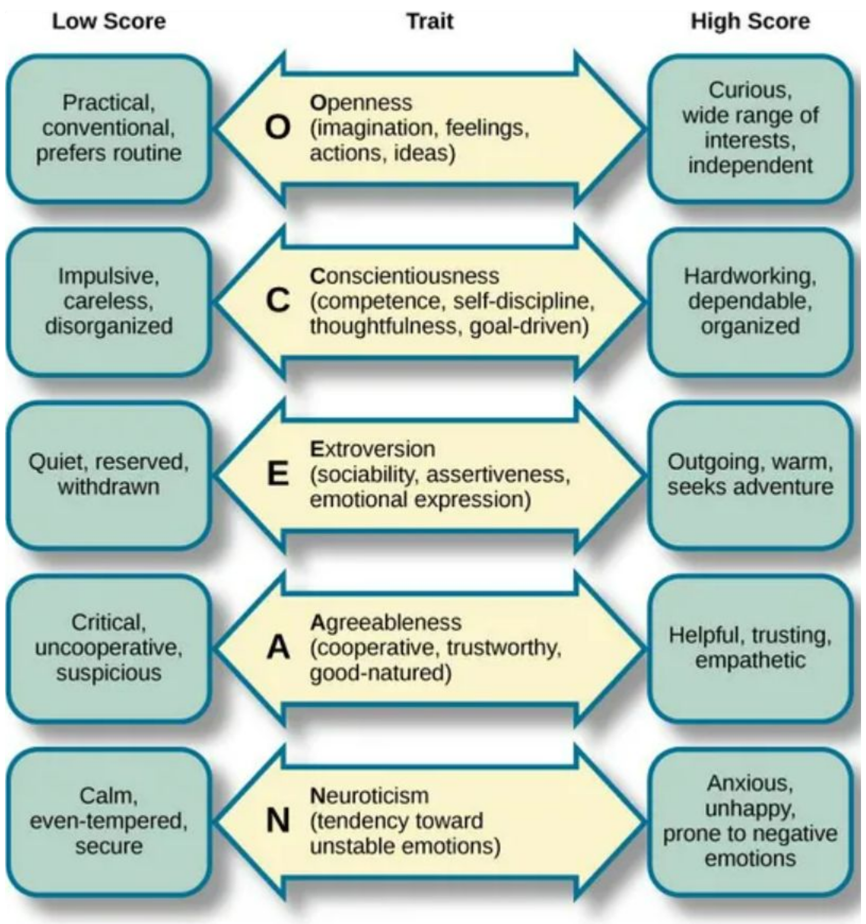
Data Description: Psychological Attributes

Personality attributes

The Big 5 Personality Traits		
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.
d	AScore	The agreeableness trait reflects individual differences in general concern for social harmony.
e	Cscore	Conscientiousness is a tendency to display self-discipline, act dutifully, and strive for achievement against measures or outside expectations.
Personality Traits Based on Risk Factor for Hazardous and Maladaptive Behaviors		
f	Impulsive	Impulsivity is a related yet distinct construct that reflects deficits in perseverance, planning, and inhibitory control
g	SS	Sensation seeking is a personality trait that reflects the tendency to pursue and enjoy novel and stimulating experiences

Data Description

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



Data Description: Drugs

Drugs	Chemical Name
Alcohol	Alcohol
Amphet	Amphetamines
Amyl	Nitrite
Benzos	Benzodiazepine
Caff	Caffeine
Cannabis	Marijuana
Choc	Chocolate
Coke	Cocaine
Crack	Crack Cocaine
Ecstasy	Ecstasy
Heroin	Heroin
Ketamine	Ketamine
Legalh	Legal Highs
LSD	LSD
Meth	Methadone
Mushroom	Magic Mushroom
Nicotine	Nicotine
Semer	Class of Fictitious Drugs Semeron (i.e control)
VSA	Class of Volatile Substance abuse

Drug excluded from the study:
Caffeine, Chocolate, Control drug - Semer

Data Description: Drugs

Drug Class Labels

7 Class labels	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

Methodology

Data Collection

- Drug consumption data set collected from UCI Machine Learning repository.

Data Processing

- Data reduction: Ignore the ID in the data
- Data transformation: create binary class for drug classes (User and Non-user) using existing class labels

Data Mining and Algorithms used

- Use of Apriori Association Mining for different drug substances consumed
- Use of Classification Models like Decision Tree and Bayes Nets for 7 class labels for every drug class *
- Similar classification was used for Binary class labels for every drug class
- Correlation of Attributes & Drugs
- Peasrson correlation - Python for correlation analysis

Measure for Evaluation

- We used Cross Validation
- Overall Accuracy
- Confusion Metrics
- And Graphs for analysing

*Best fit Classification model to be selected among Decision Tree, Neural Network and Bayes Nets

Analysis: Association

Problem Statement

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.

This problem can be addressed by performing Market Basket analysis with the use of the Apriori algorithm. Market Basket Analysis (MBA)is a data mining technique that allows for identification of associations between attributes. In this case, it the tool was used to identify patterns of consumption between the different drugs classes and find strong association rules between drugs that have deadly and highly dangerous interactions.

Minimum support: 0.2 (377 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 16

Best rules found:

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)

Analysis: Association for Synthetic Drugs

Minimum support: 0.2 (377 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 16

Best rules found:

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).

Analysis: Without Alcohol

Minimum support: 0.1 (188 instances)

Minimum metric <confidence>: 0.7

Number of cycles performed: 18

Best rules found:

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)

Insights

- 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.
- Synthetic drugs show reincidence in alcohol consumption. And they also show lethal combinations like Benzo, cocaine and alcohol.
- “Speedballing” practices can be identified.
- Alcohol is commonly combined with cocaine, which creates a lethal combination in the liver called cocaethylene.
- 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.
- Strong associations between stimulants, which are known to put high strain on the heart.
- When alcohol was not considered, lift values much greater than 1.0 were found. Which indicates that transactions containing drug A tend to contain drug B more often than transactions that do not contain drug A.

Analysis: Classification

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 drug.

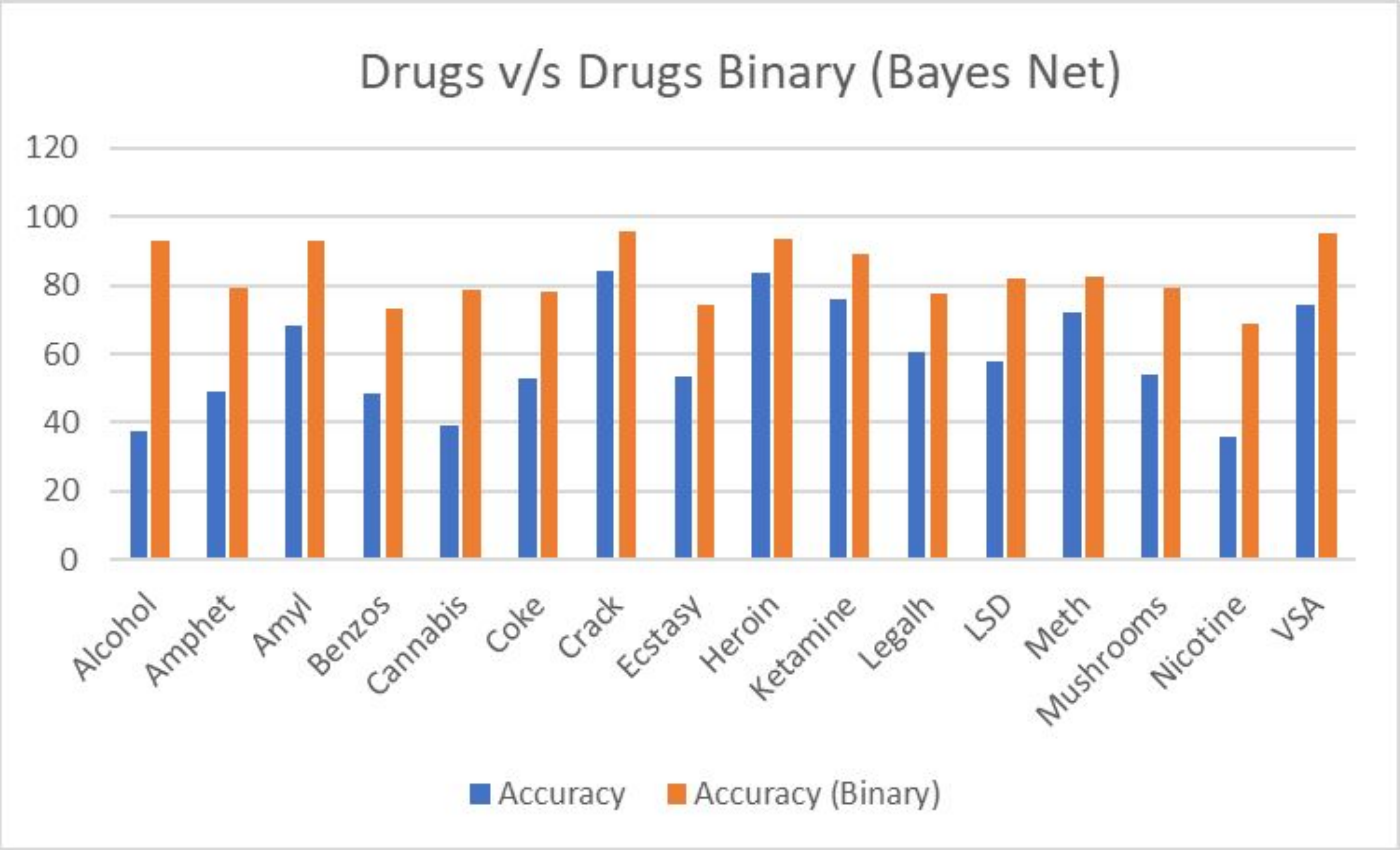
Parameters:

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

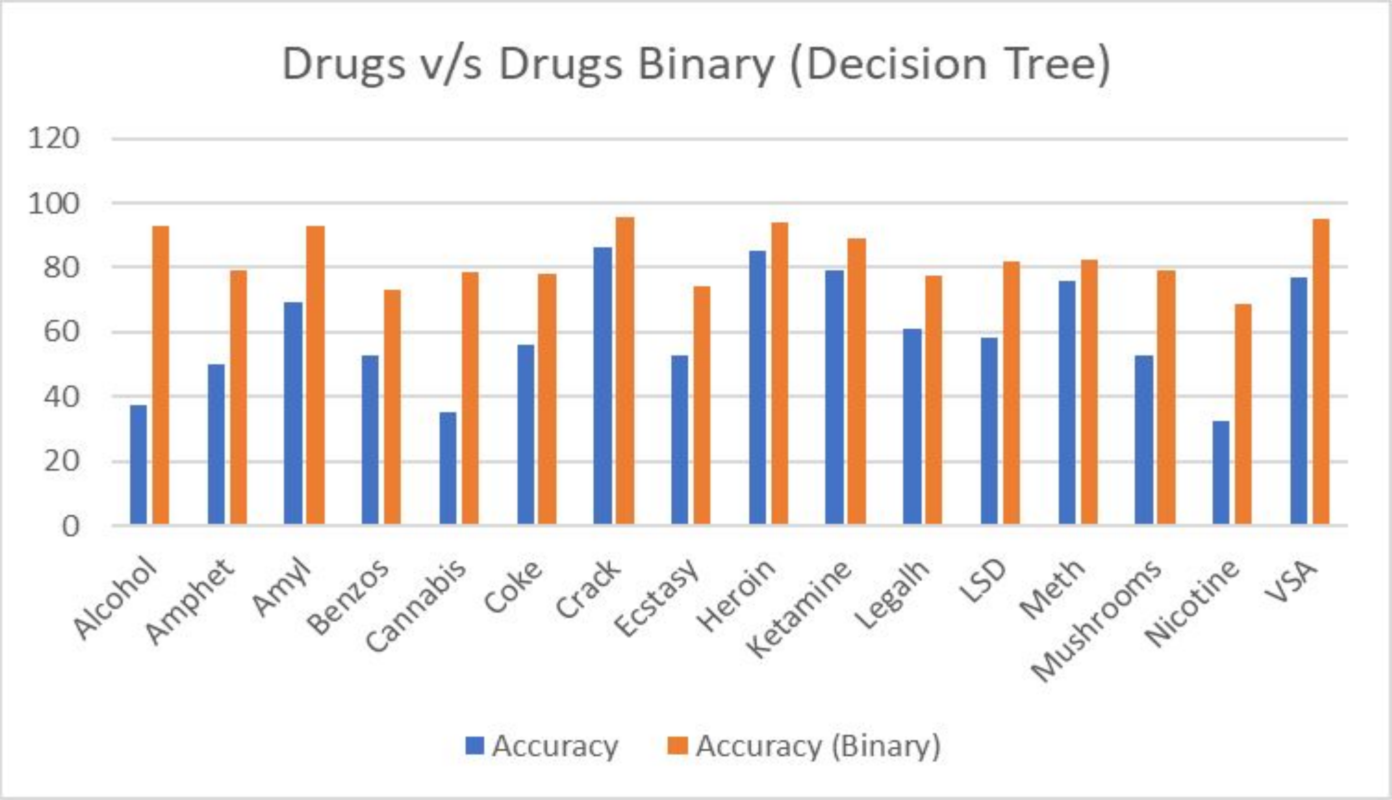
Analysis: Classification

Drug Class Labels	
7 Class labels	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
Binary Class Labels	**Non User:**
	CL0 Never Used
	CL1 Used over a Decade Ago
	CL2 Used in Last Decade
	User:
	CL3 Used in Last Year
	CL4 Used in Last Month
	CL5 Used in Last Week
	CL6 Used in Last Day

Bayes Net Analysis



Decision Tree Analysis



Observations

Confusion matrix

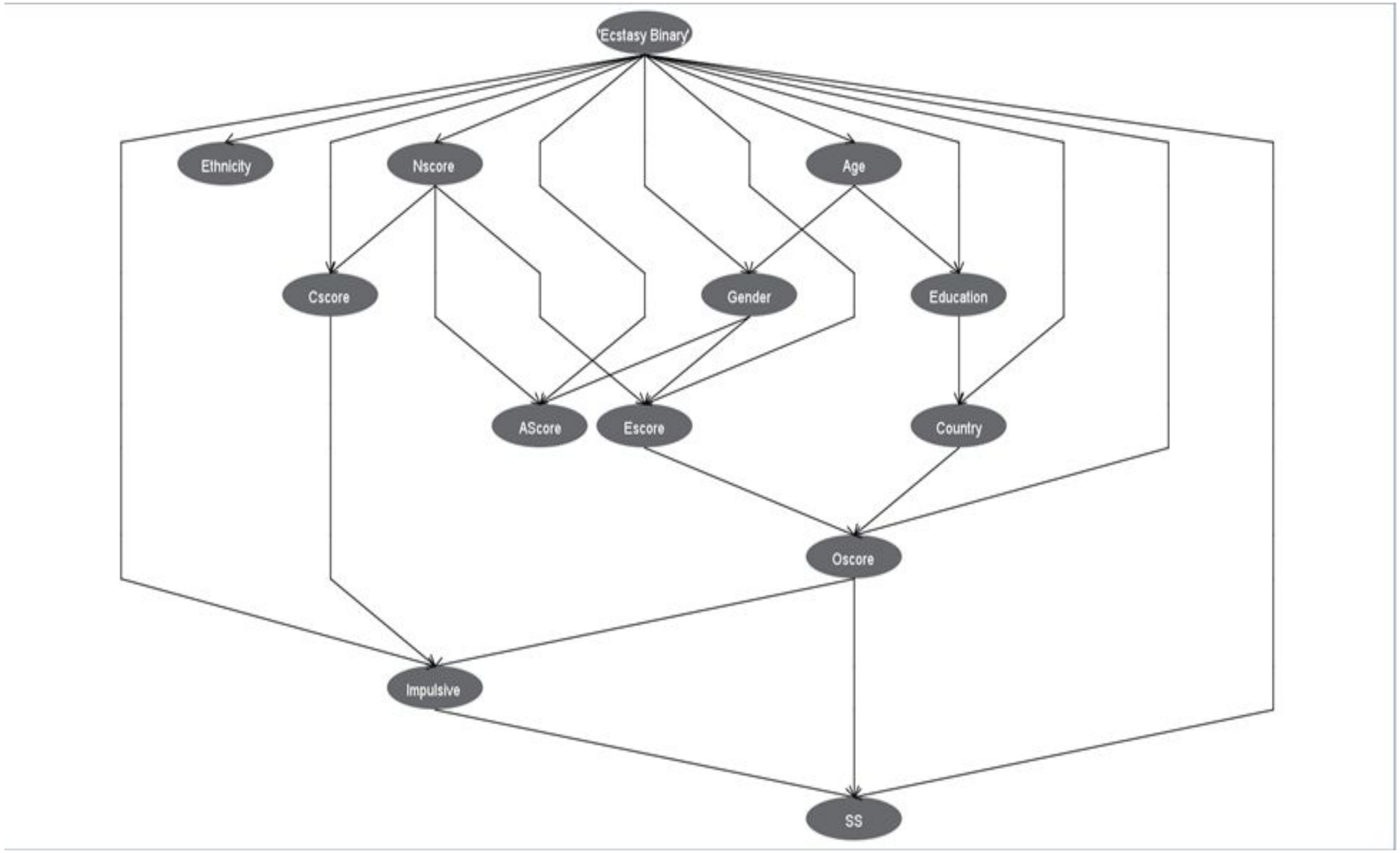
Drugs

```
  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
```

Drugs Binary

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

Visualising graph: Bayes Net



Probability Distribution Table For Age

'Ecstasy Binary'	25-34	35-44	18-24	65+	45-54	55-64
0	0.259	0.227	0.232	0.014	0.201	0.066
1	0.243	0.086	0.626	0.001	0.037	0.007

Analysis: Correlation Analysis

3. Identify if there is any correlation between

- Demographic attributes (e.g. age, education, ethnicity, etc.) and drug substance consumption
- Behavioral attributes (e.g. Escore, Ascore, Impulsive Score, etc.) and drug substance consumption
- Correlation within Behavioral attributes

Predictability of binary drug classes using demographic and behavioral attributes separately

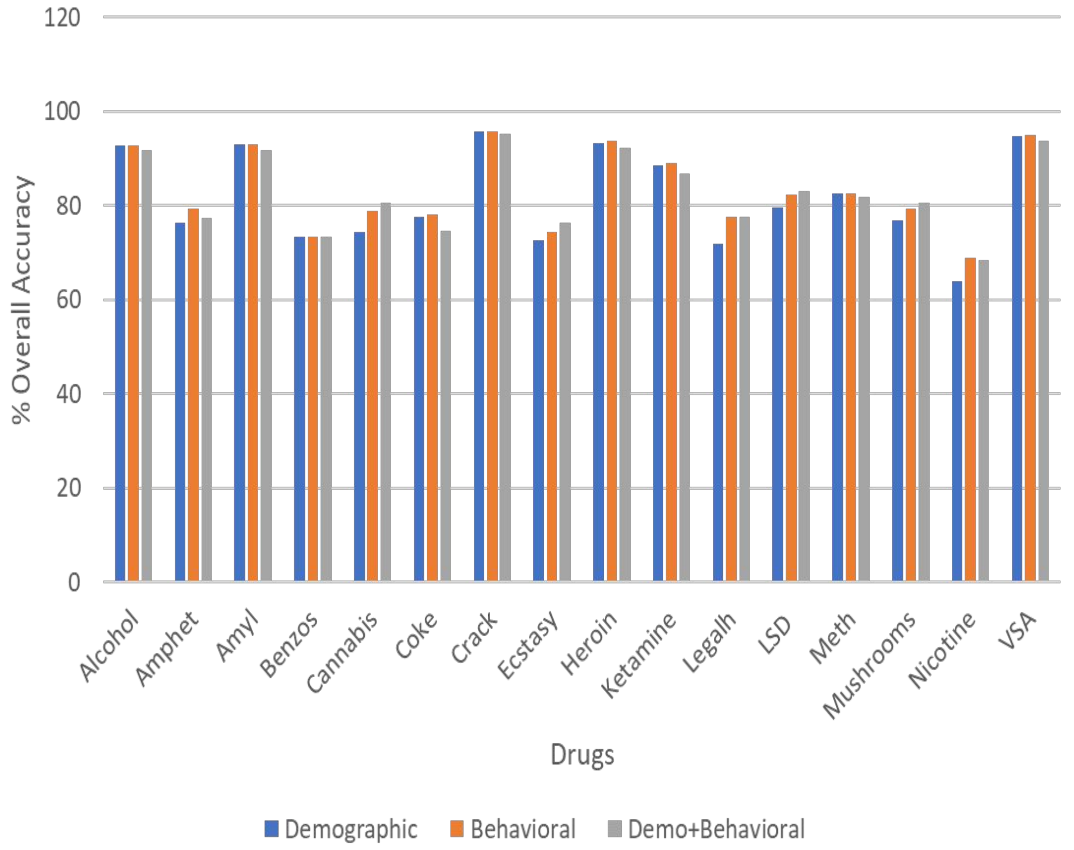
Method Used: Classification Performance using Bayes Net

Parameters: 5 -fold cross-validation, max number of parent = 5

Observations:

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

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



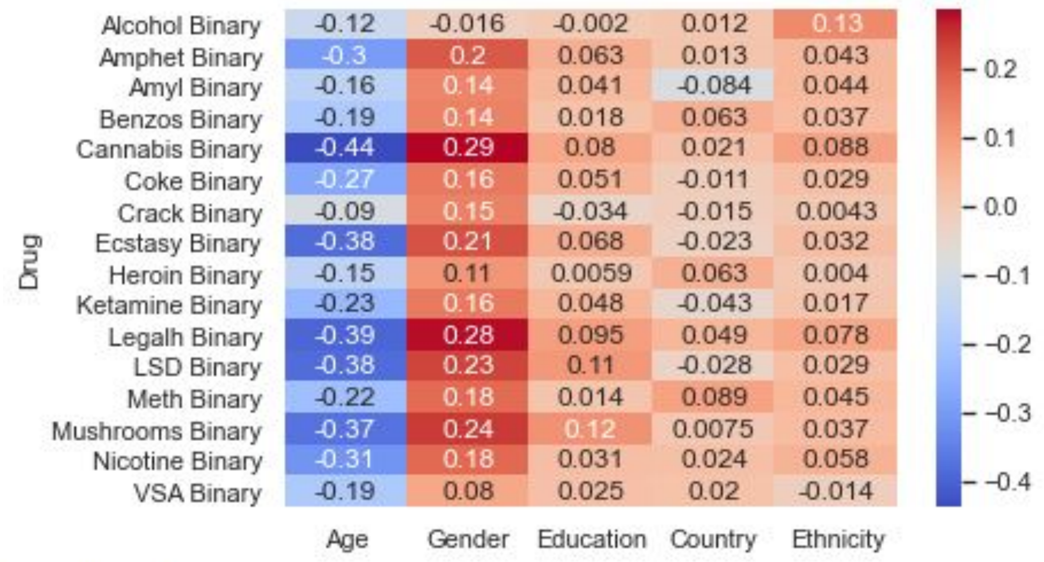
Correlation of each demographic features and binary drug classes

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

Observations:

1. Education, Country, Ethnicity show no significant correlation
2. Males are showing a higher correlation to drug consumption
3. Lower age groups have a higher correlation to drug consumption

Heat Map of Correlation



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

Cannabis Binary	0	1
Gender		
F	0.61	0.39
M	0.33	0.67

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

Correlation of each behavioral feature and binary drug classes

Method used: Using Corr(correlation) function in Pandas using Heatmap

Observations:

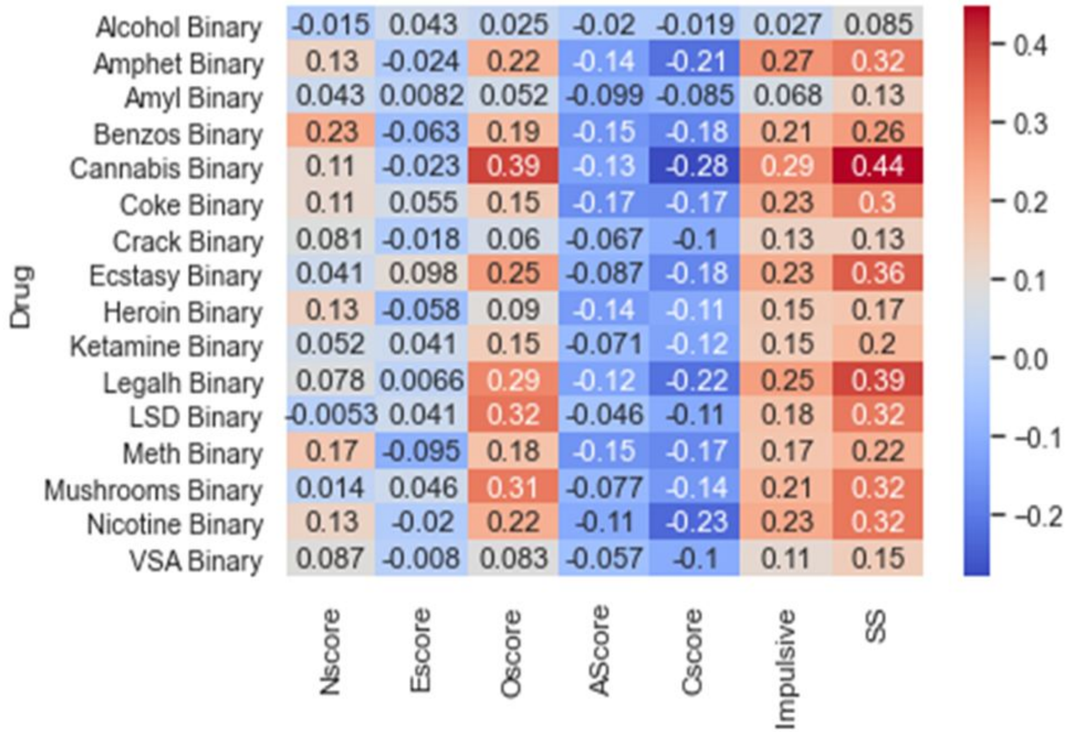
1.Cannabis is the only drug that shows moderate correlation with behavioral attributes

a. Higher Oscore (Openness to trying new things) and SS score (Sensation seeking) have shown to have positive correlation
b. Cscore (Conscientiousness – Self-discipline) is negatively correlated to Cannabis

2.Ecstasy and Legalh have also shown some positive correlation with SS score (Sensation seeking)

3.Nicotine has shown some negative correlation with Csore (Conscientiousness – Self-discipline)

In general, Psychotropic drugs have some correlation with behavioral attributes



Correlation within Behavioral Features

Method Used: Pearson correlation in Pandas using heatmap

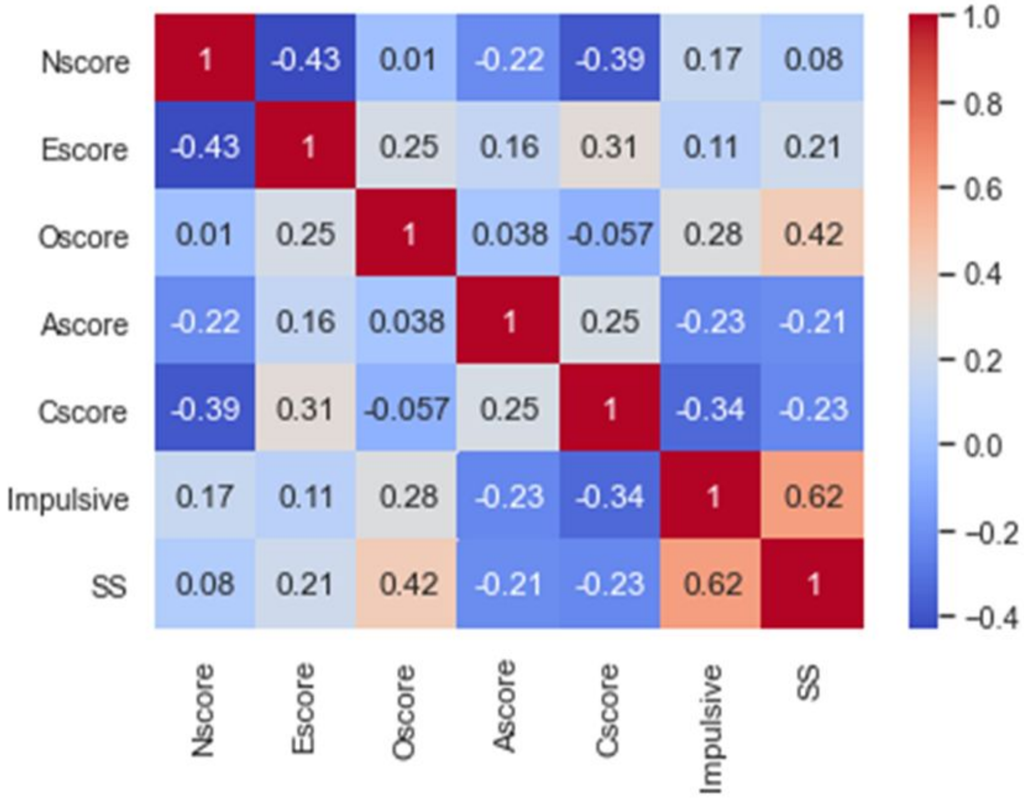
Observations:

1.SS score (Sensation seeking) has shown a positive correlation with Impulsive score and Oscore (Openness)

2.Nscore (Neuroticism – e.g Anxitey, depression) has shown a negative correlation with Escore (Extraversion) and Cscore (Conscientiousness)

3.Impulsive score and Cscore (Conscientiousness) negatively correlated

In general, observed correlations validate known assumptions about behavioral traits



Conclusions

- Data mining techniques were proven useful at building several models that allowed to predict behavior and classify instances of drug consumption.
- Market Basket Analysis allowed to discover hidden patterns within the data set, and allowed to establish relationships between the attributes with high levels of confidence.
- Binary classification of drugs showed better accuracy than drugs.
- Bayes Net classification model helped in better visualization and identifying attributes dependencies for drugs consumption.
- Psychotropic drugs have some correlation with behavioral attributes
- Drugs 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
- 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)

References

- E.Fehrman,A.K.Muhammad,E.M.Mirkes,V.EganandA.N.Gorban,"TheFiveFactorModelofpersonalityan devaluationofdrugconsumptionrisk.,"arXiv[WebLink],2015
- Datasource: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).TheU.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
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2678841/#:~:text=Finally%2C%20the%20ImpSS%20scale%20is,range%20from%200%20to%2019.

thank you!