Machine Learning & Data Mining
Project 1 - Data: Feature extraction, and visualization

Group 72: Axel Månson Lokrantz, Isak Wilkens, Julius Ekberg Bretto

2023-08-05

Contribution

	Section 1	Section 2	Section 3	Section 4	Exam questions
s232081	30%	33%	40%	30%	33%
s232082	40%	33%	30%	30%	33%
s231401	30%	33%	30%	40%	33%

Exam Problems

Question 1

Answer: D
We believe that the correct option for this question is D. Since we consider D correct we deem the other options incorrect by induction.

x1 (Time of day): we consider this interval since distance between the objects can be measured in the 30 minute intervals. They can also be manipulated with addition/-subtraction. We discussed whether this was interval or ordinal, considering the situation where a camera records the traffic; we can then add or subtract intervals to move back and forth in the recording.
x2 (Traffic lights): is ratio, because there is a clear zero point (0 broken traffic lights) within the category.
x7 (Running over): is ratio, following the same argument as above regarding the broken traffic lights – there is a clear zero point, 0 run over incidents.
y (Congestion level): is ordinal since it is a ranking and can be ordered but we do not know the distance and it has no clear zero point.

Question 2

Answer: A
The p-norm distance is equal to 7, hence answer A is correct. This was achieved by using the formula for calculating the max-norm distance of a n-dimensional vector. We constructed a program in R to execute the calculations:

```
# Create vectors x1 and x2
x1 <- c(26, 0, 2, 0,  0, 0, 0)
x2 <- c(19, 0, 0, 0, 0, 0, 0)

# Calculate the p-norm for p=infinity
inf_norm <- max(abs(x1 - x2))
print(paste("Infinity norm: ", inf_norm))
```

Following this formula, it is evident that the remainder of alternatives, B, C, D and E, are incorrect.

Question 4

Answer: D
The matrix V contains the eigeinvalues which can be seen as loadings in the SVD decomposition. The given observation options show either high or low values in an attribute. If the high/lows in the observation are similar values to a principal component in the matrix the impact will have a positive value onto the projection onto it and vice versa. Given this, observation D is the only correct answer since its values of high/low are similar to the loadings of principal component 2 and is stated to have a positive impact.

Question 5

Answer: C
In order to calculate the Jaccard similarity, we need to first calculate the intersection and the union of the problem set. Immediately we can distinguish the words "the" and "words" in both of the text strings s1 and s2, giving us an intersection value of 2 elements. Furthermore, the union would – in the case of not having encoded the strings from a total vocabulary of the size M = 20000 – been equal to the total amount of words in both of the text strings s1 and s2 minus the amount of the intersection (in this case the value of 13). However, since the problem formulation includes the information that the text string s1 and s2 have been, through a BoW-encoding, created from a vocabulary of M = 20000 words, we interpreted the union of being equal to 20000. Hence the answer, according to the formula of Jaccard Similarity = Intersection / Union, is 2/20000=0.0001, answer C.

We calculated this through constructing a Python-script:

```
s1 = "the bag of words representation becomes less parsimoneous"
s2 = "if we do not stem the words"
```

```
words_s1 = set(s1.split())
words_s2 = set(s2.split())

intersection = len(words_s1.intersection(words_s2))
union = 20000
jaccard = intersection / union
print(jaccard)
```

1. Description of the Data Set

1.1 Data Overview

The data set contains records for 1885 respondents for which 12 attributes are known. The attributes cover personality measurements, demographic characteristics and data related to the respondents consumption of 18 central nervous system psychoactive drugs. The main problem of interest is to understand how personality traits and personal background relate to the consumption of drugs. Personal measurements are NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity. The 18 drugs measured in the study are: alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) introduced to identify over-claimers.

1.2 Link to Data Set

Link

1.3 Summary of Data Set

The dataset was collected from an online survey by A.K. Muhammad, E.M. Mirkes, V. Egan, E. Fehrman, and A.N. Gorban and detailed in a research paper titled 'The Five Factor Model of Personality and Evaluation of Drug Consumption Risk.' The paper's primary objective was to investigate the correlations between personality traits and drug consumption around the world. It aimed to address classification problems associated with personality traits and the risk of drug consumption. The study found that there were three distinct groups of drugs with strongly correlated consumption patterns named after their central drug: ecstasy, heron and benzodiazepines. High quality of classification was reach with sensitivity and specificity being grater than 70%. The authors achieved best results for the drugs cannabis, crack, ecstasy, legal highs, LSD and VSA. All input attributes are categorial and are quantified.

1.4 Attributes for Classification and Regression

By analyzing the data on drug consumption the main machine learning aim is to identify patterns based on the respondents answers. Personality traits and personal background attributes can help predict the likeness of drug use.

As a regression task a risk estimation model could be designed to estimate an individuals risk level for drug use based on their personality traits and demographic information. This type of information can be valuable for prevention efforts or public health initiatives. Regression could also be used to predict frequency or quantity of drug consumption.

For the classification problem, clustering techniques could be utilized to group drugs into clusters. Subsequently, a classification task could predict which drugs individuals are more likely to use in conjunction. For instance, correlations between hallucinogenic drugs like LSD and mushrooms could be explored.

Using this data for a classification clustering techniques could be used to group individuals into clusters where a classification task can predict cluster membership. This can help identify patterns amongst drug users.

Before applying classification and regression some data manipulation might be necessary. The be able to visualize and perform statistics on the data it will be mean centered. The data from the study is mostly quantified but the drug use levels are estimated in a different scale which it might be helpful to transform all of the data into similar scales.

2. Detailed Explanation of the Data Set

2.1 Attribute Data

See appendix table 1 for a detailed description of the data attributes where they are categorized discrete/continuous, Nominal/Ordinal/Interval/Ratio. The responses of the first person with id 1 is also shown there as an example.

2.2 Missing values and corrupted data

There are no reports of missing values or corrupted data in the data description.

2.3 Basic summary statistics of the attributes.

Some summary statistics for the data is shown in appendix table 2, since the data is quantified their real world representation is critical to keep in mind working with the data. The standard deviation column given an idea about the spread of values around the mean. For instance "education" has a high standard deviation of 0.9501 indicating a wide spread of education levels. For the median values, if the value is different from the mean it might suggest that the data is skewed. For example "impulsiveness" has a mean of 0.0072 but a median of -0.2171. The range column shows the difference between the maximum and minimum value in the data. For example "country" has a range of 1.5309 suggesting a large variation in this attribute. We can see some representative features in the data, attributes with higher means like "Caffeine", "Chocolate" and "Alcohol" could be more representative due to their higher usage rates.

3. Data Visualizations

3.1 Outliers

An outlier is an observation or data point that significantly deviates from the rest of the data in a dataset. In Figure 3, outliers can be observed in the attributes 'Semer' and 'Ethnicity.' 'Semer' is a fictional drug that does not exist; therefore, respondents who indicated that they had used Semer were removed from the dataset. In the 'Ethnicity' attribute, an outlier is present due to the fact that only 0.16% of the respondents belonged to the Mixed-Black/Asian category, while 91.25% belonged to the white category.

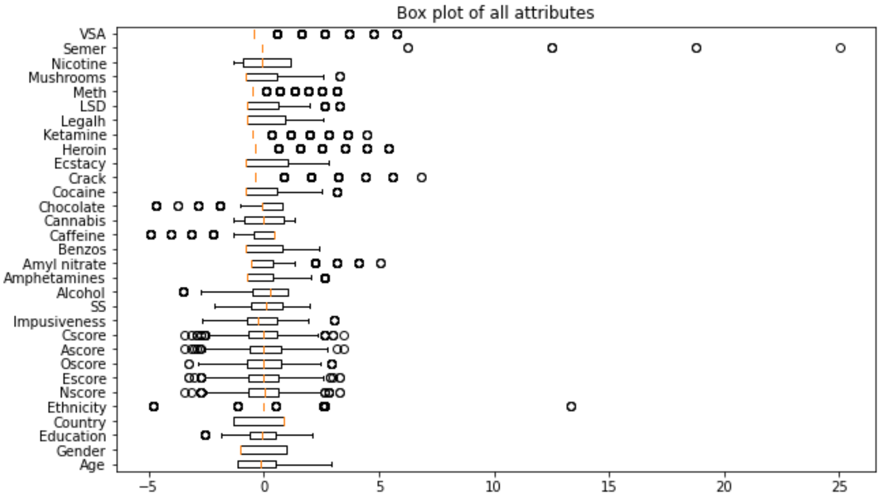

Figure 1: Box plot of outliers in the data set.

3.2 Normal distribution of attributes

Normal distribution, often referred to as a bell curve, indicates that the data is symmetric and centered around the mean, where the right and left halves of the distribution are mirror images of each other. Based on visual inspection of histograms, the group has concluded that none of the attributes appear to follow a perfect normal distribution. However, below, the group has plotted all the attributes and through visual inspection attributes 6 to 10 exhibit distributions closest to the normal.

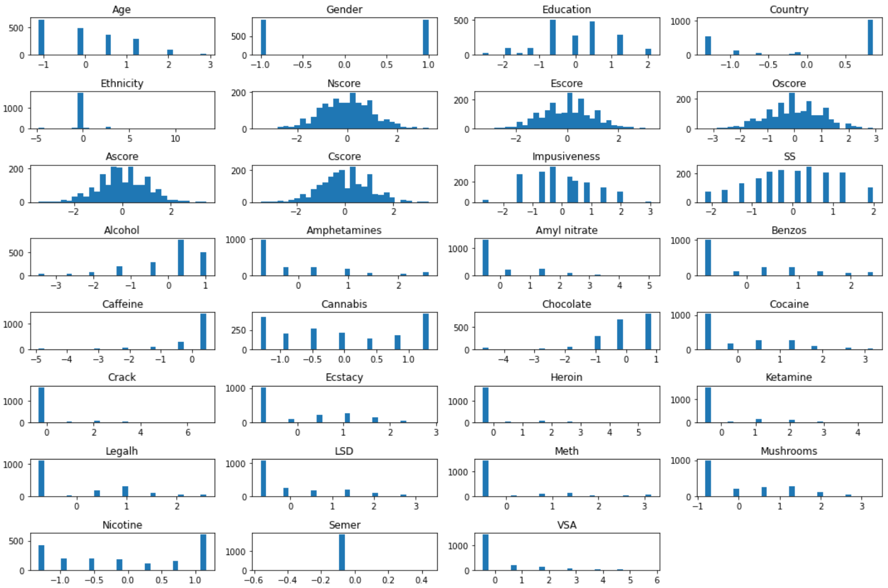

Figure 2: Some of the attributes that come closest to normal distribution.

3.3 Correlated variables

To address the issue of having different scales for our data the group decided to mean center it by calculating the mean value of each attribute and then subtract it from all attribute values. Through Z-score normalization the data was then standardized. The Z-scores, represent how many standard deviations an observation is from the mean for each attribute. It helps when comparing and analyzing data on the same scale and can be useful for identifying outliers.

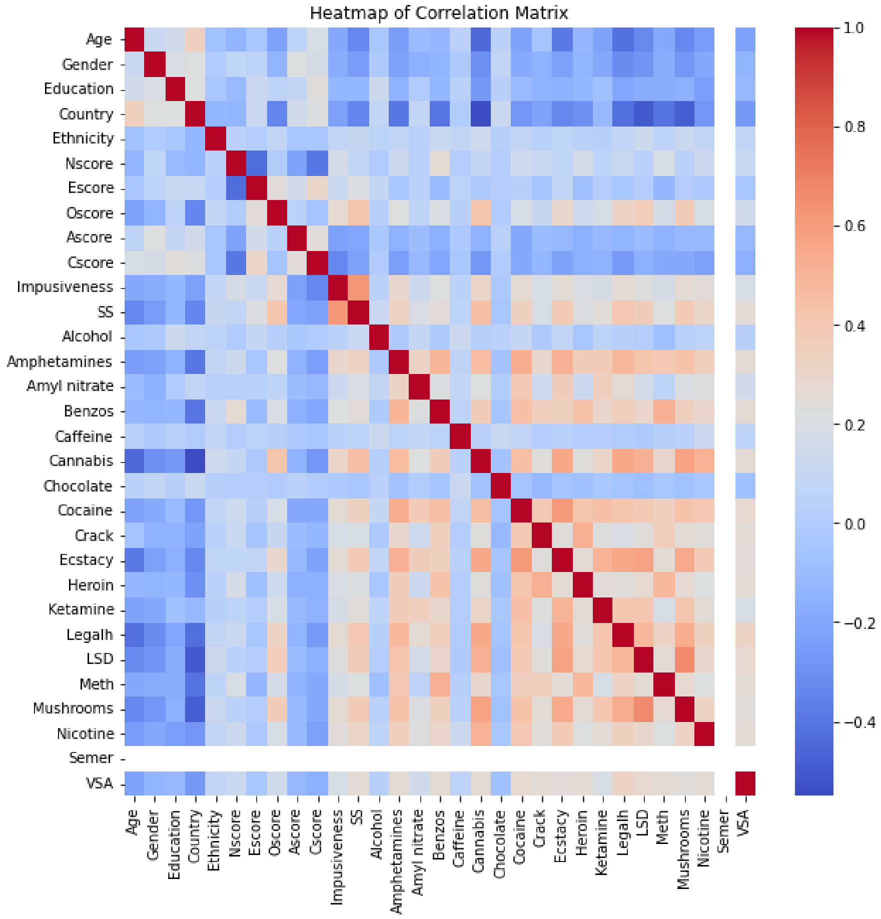

Figure 3: Heat map representation of all attributes in the data set.

The heatmap visually represents the correlations between attributes, where the colors indicate the strength and direction of the correlations.

The heatmap shows how strongly and in what direction (positive or negative) different attributes in the dataset are correlated. Red indicates strong correlation while blue indicates the opposite. The red diagonal line that runs across the image shows how each and every attribute has a 1.0 correlation coefficent with itself.

3.4 Machine Modelling Aim

The data that has been gathered does not have many clear faults and few values that are outliers. Looking at the correlation heat-map we can see a high correlation between the different types of drugs but a lower correlation between the drugs and the personal information. Some of the data does not follow this trend and might lead to interesting conclusions that can be used in our main machine learning aim. If the evidence of drug usage has too few meaningful correlations with the other attributes, modelling could instead be done around drug clustering and seeing how these clusters adhere to the rest of the data. The researchers conducting the original study found some clusters of drugs where there was a main drug defining the cluster.

3.5 Variation PCA

Principal Component Analysis or PCA, is a dimensionality reduction technique that transforms the data such that the greatest variance by any projection of the data comes to lie on the first principal component, the second greatest variance on the second principal component and so forth. In some datasets, just a few principal components might capture a large portion of the total variance which indicate that the original attributes are highly correlated. In other datasets, such as in the case of our data set, the variance might be spread out over many principal components which indicates that the attributes are not highly correlated. To cover approximately 90% of the total variance of the data set 23 principal components are needed. As seen in figure 4, the curve is relatively smooth, without any noticeable inflection points or 'elbows'.

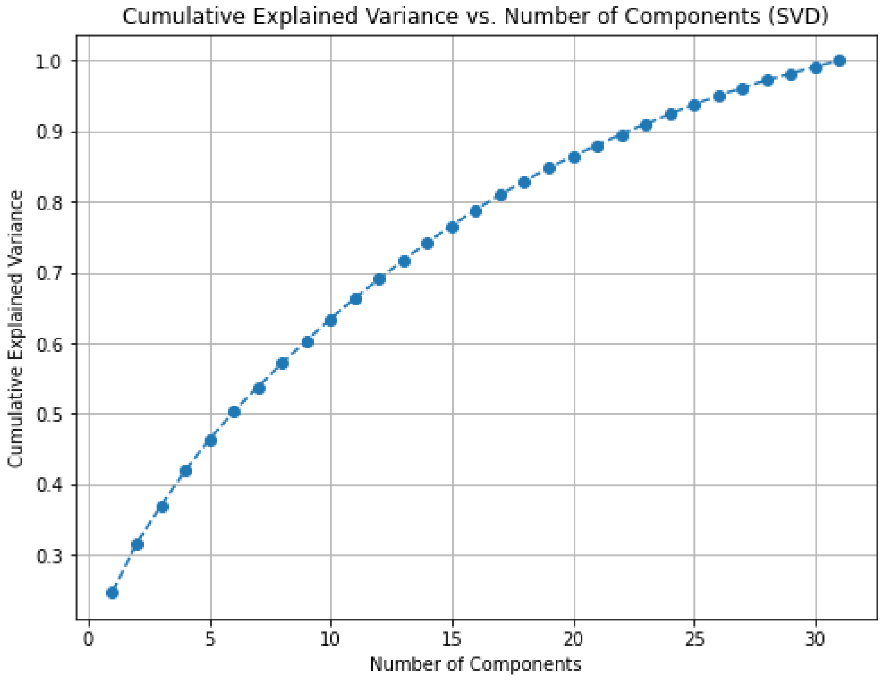

Figure 4: Cumulative explained variance vs. number of principal components.

3.6 PCA Component Coefficients

As previously mentioned, to cover around 90% of the total variance of the data set, 23 principal components are required. However, plotting the PCA Component Coefficients in 23 different directions gets very messy and makes it hard to distinguish any valuable information. In figure 5, we made the decision to only plot 3 principal components for clarity.

The different colors (red, green and blue) denote different principal components, meaning different directions or axes in the third dimension along which our data is projected. The PCA component coefficients indicate the direction and strength of the relationship between each attribute and the principal component. The larger the absolute value of the coefficient, the more important is the specific attribute. For example, for PC0, the Cannabis-attribute has a coefficient of 3 while the Age-attribute has a coefficient of around -2. This implies that when observing data objects (individuals), an increase in PC0 would suggest a tendency toward more frequent cannabis use, while a decrease in PC0 would suggest a tendency toward lower ages.

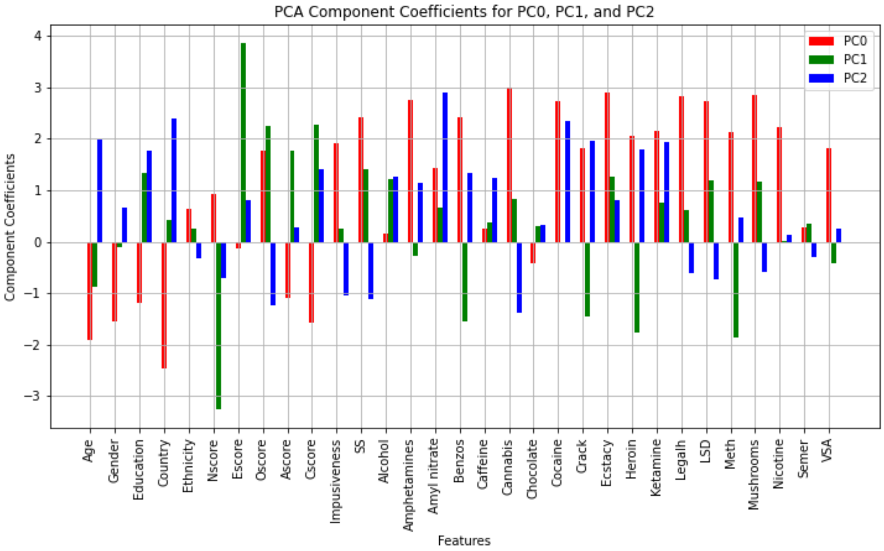

Figure 5: Principal component coefficents for PC0, PC1 and PC2.

3.7 PCA Data Projection

To further visualize the data we can create a 2D representation of our high dimensional data. This can help observe patterns or outliers in the data. This is done by projecting the data onto the two principal components that cover the largest amount of variance. This is done in Figure 6 where we can see that the spread along PC0 is wider and therefore explains more of the variance in the data. It is hard to identify any clear clusters in this representation of the data. We can see that some areas have a high density which tells us that many of the attributes in the data are similar. This corresponds with the type of data that is used where many of the drugs and personality traits are similar types of data.

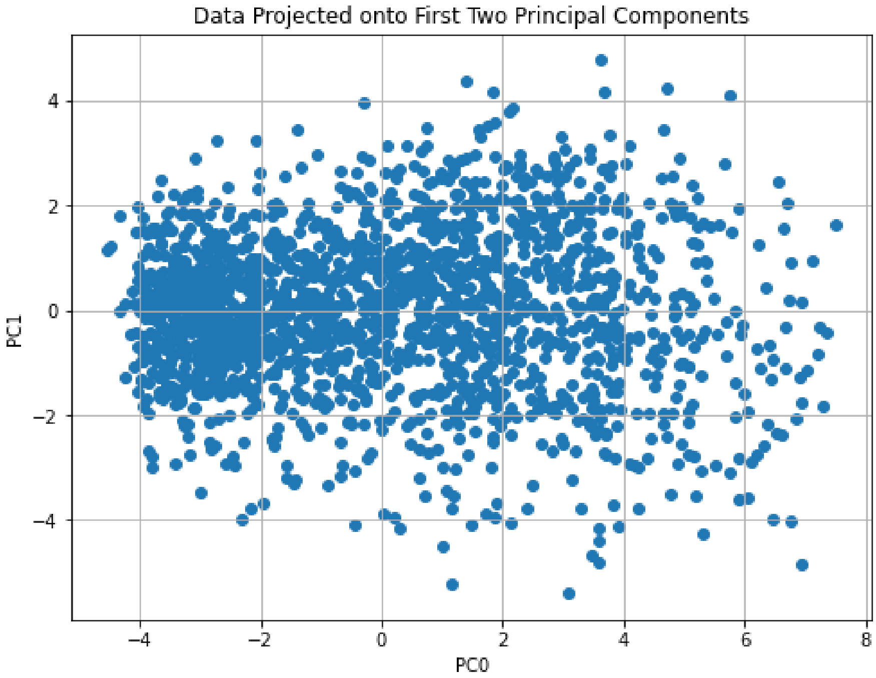

Figure 6: Data projected onto first two principal components.

4. Concluding Discussion

The initial points of interest in choosing this particular data set to work with was to analyze and understand how different personality traits and demographics relate to the consumption of drugs and if there exists any correlation between these factors. More specifically, we wanted to construct a machine learning model that would be able to distinguish what types of drugs a person has used or the probability that they will use it, as well as their frequency of use, depending on their personality traits (using the NEO-FFI-R, ImpSS and BIS-11 models) as well as their demographics (level of education, age, country of residence, ethnicity, gender). The aim is that the model will be able to identify drug consumption patterns based on these factors.

Firstly, regarding outliers, we made the decision to completely remove the data objects (individuals) that stated they had used the fictitious drug Semeron. This was because we deemed their responses were probably not accurate, since ”over-claiming” individuals tend to exaggerate. Furthermore, the Ethnicity category Mixed-Black/Asian was also deemed an outlier, but in this case we decided to keep these data objects in the data set. This choice was based on the fact that they were classified as outliers primarily due to their small representation among the respondents, and this unique categorization doesn’t affect our analysis of drug use and personality traits in (any and all) individuals.

On that note, the data set shows a very apparent skewing when it comes to ethnicity. 91.25% of the respondents identify as white. The dataset’s objective is to depict drug consumption patterns based on personality data for individuals in a general context, encompassing all ethnic backgrounds rather than being limited to one group (ethnic whites). To avoid introducing a bias in the dataset for ethnicity, or any other attribute for that matter, the data set should be distributed evenly across various ethnic groups within the respondent population. Currently, the dataset does not reflect a representative sample that mirrors the global demographic mean. The same can be said when discussing the country of residence of most the respondents, since a) all of the respondents live in English-speaking countries of the West and b) even here, the distribution of respondents is skewed, with a vast majority living in the UK, and the runner-up being the USA.

Furthermore, the Heat Map (Figure 3, 3.3 Correlated variables, p. 7) is a very interesting data visualization. The Heat Map represents the paired correlation between any and all attributes in the data set. From a quick glance, it's easy to distinguish a very strong correlation between the usage of the hallucinogenic drug LSD and mushrooms (with hallucinogenic qualities). This entails that respondents that do use LSD very often use mushrooms as well, and vice versa, which makes sense since the sensory effects are not that dissimilar. Conversely, Chocolate and Caffeine have low to medium scores across every category, meaning that there is no specific correlation between those two drugs and other drugs or personality demographics/traits.

Regarding the personality traits, having a high score of the attribute "Impulsiveness" made a person, unsurprisingly, more prone to drug consumption, as can also be deduced from the heat map. The same can be said for the attribute "SS" (Sensation Seeking) – this was the personality trait that scored the overall highest across all drugs, indicating that people with this personality trait are more prone to all drug use. For some drugs the attribute "Oscore" (measuring Openness) also showed on higher correlation, which was more or less expected. Lastly, the attribute "Nscore" measuring tendencies toward negative feelings like anxiety, depression and self-doubt are relatively strongly correlated to the use of Benzodiazepines, Meth and Heroin.

What is interesting, as previously briefly mentioned, is that there appears to be an overall strong correlation between drugs, but not as strong a correlation between personality traits/demographics and drugs. This may mean that trying a drug, any drug, might make an individual more prone to drugs than any personality trait or demographic can, solidifying the concept of "gateway drugs". Similarly, some of the data does not follow this trend, and we are therefore remaining positive that we can achieve our machine learning aim of creating a model that can assess the risk of drug consumption based on demographic and personal data. If this proves too far-fetched, there is always the opportunity of clustering data and analyzing how these different clusters relate to each other. For example, if one drug in the drug cluster defining the whole cluster, and so on.

Conclusively, we have made major advancements in structuring, analyzing and understanding the data presented. We have distinguished many interesting correlations, both positive and negative, and plan to use these in the future when constructing our machine learning model. We, as a group, remain positive that we can achieve our aim of a risk estimation model for drug consumption (as well as predicting frequency of drug consumption) with basis in an individual's personality traits and demographics.

Appendix

Table 1

Table 1: Attribute Data

Attribute	NumType	Type	Example data
ID	Discrete	Nominal	1
Age	Continuous	Ordinal	0.49788
Gender	Discrete	Nominal	0.48246
Education	Continuous	Ordinal	-0.05921
Country	Discrete	Nominal	0.96082
Ethnicity	Discrete	Nominal	0.12600
Nscore	Continuous	Ratio	0.31287
Escore	Continuous	Ratio	-0.57545
Oscore	Continuous	Ratio	-0.58331
Ascore	Continuous	Ratio	-0.91699
Cscore	Continuous	Ratio	-0.00665
Impulsive	Continuous	Ordinal	-0.21712
SS	Continuous	Ordinal	-1.18084
Alcohol	Discrete	Nominal	CL5
Amphetamines	Discrete	Nominal	CL2
Amyl Nitrate	Discrete	Nominal	CL0
Benzos	Discrete	Nominal	CL2
Caffeine	Discrete	Nominal	CL6
Cannabis	Discrete	Nominal	CL0
Chocolate	Discrete	Nominal	CL5
Cocaine	Discrete	Nominal	CL0
Crack	Discrete	Nominal	CL0
Ecstasy	Discrete	Nominal	CL0
Heroin	Discrete	Nominal	CL0
Ketamine	Discrete	Nominal	CL0
Legal Highs	Discrete	Nominal	CL0
LSD	Discrete	Nominal	CL0
Methadone	Discrete	Nominal	CL0
Mushrooms	Discrete	Nominal	CL0
Nicotine	Discrete	Nominal	CL2
Semer	Discrete	Nominal	CL0
VSA	Discrete	Nominal	CL0

Table 2

Table 2: Summary Statistics for Attributes

Attribute	Mean	Standard Deviation	Median	Range
Age	0.0346	0.8784	-0.0785	3.5437
Gender	-0.0003	0.4826	-0.4825	0.9649
Education	-0.0038	0.9501	-0.0592	4.4203
Country	0.3555	0.7003	0.9608	1.5309
Ethnicity	-0.3096	0.1662	-0.3169	3.0143
Nscore	0.0000	0.9981	0.0426	6.7383
Escore	-0.0002	0.9974	0.0033	6.5479
Oscore	-0.0005	0.9962	-0.0193	6.1755
Ascore	-0.0002	0.9974	-0.0173	6.9287
Cscore	-0.0004	0.9975	-0.0067	6.9287
Impulsiveness	0.0072	0.9544	-0.2171	5.4568
SS	-0.0033	0.9637	0.0799	4.0002
Alcohol	4.6350	1.3313	5.0000	6.0000
Amphetamines	1.3406	1.7836	0.0000	6.0000
Amyl nitrate	0.6069	1.0642	0.0000	6.0000
Benzos	1.4653	1.8673	0.0000	6.0000
Caffeine	5.4838	1.1146	6.0000	6.0000
Cannabis	2.9894	2.2874	3.0000	6.0000
Chocolate	5.1066	1.0893	5.0000	6.0000
Cocaine	1.1613	1.5130	0.0000	6.0000
Crack	0.2976	0.8371	0.0000	6.0000
Ecstasy	1.3141	1.6476	0.0000	6.0000
Heroin	0.3740	1.0348	0.0000	6.0000
Ketamine	0.5692	1.2200	0.0000	6.0000
Legalh	1.3560	1.7896	0.0000	6.0000
LSD	1.0615	1.4911	0.0000	6.0000
Meth	0.8265	1.6466	0.0000	6.0000
Mushrooms	1.1873	1.4663	0.0000	6.0000
Nicotine	3.2005	2.4139	3.0000	6.0000
Semer	0.0095	0.1593	0.0000	4.0000
VSA	0.4334	0.9624	0.0000	6.0000