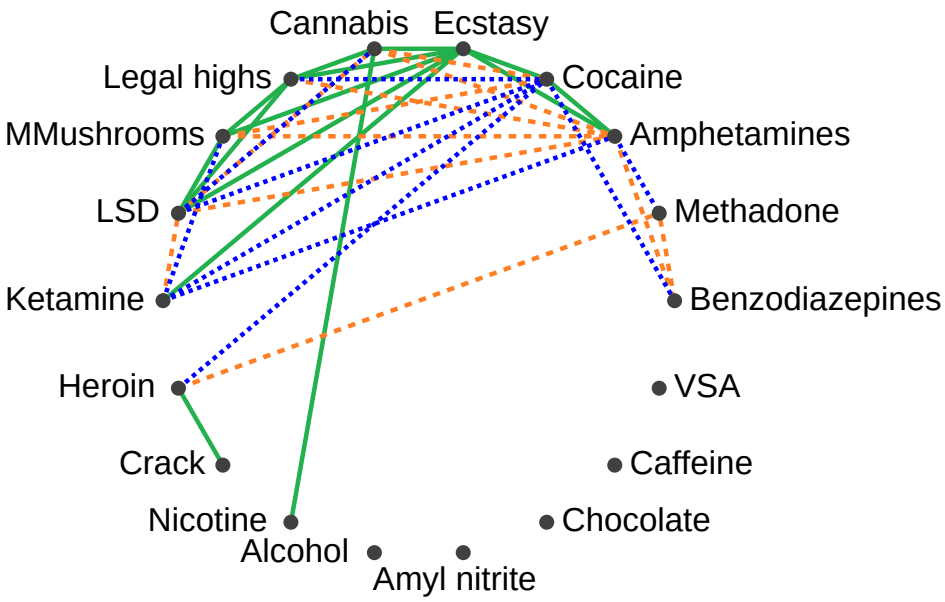
UNIVERSITY OF LEICESTER



Drug consumption, collected online in 2011-2012

Elaine Fehrman
Men's Personality Disorder and National Women's Directorate, Rampton Hospital, Retford, Nottinghamshire, DN22 0PD, UK, Elaine.Fehrman@nottshc.nhs.uk

Vincent Egan
Department of Psychiatry and Applied Psychology, University of Nottingham, Nottingham, NG8 1BB, UK, Vincent.Egan@nottingham.ac.uk

Evgeny M. Mirkes
Department of Mathematics, University of Leicester, Leicester, LE1 7RH, UK, em322@le.ac.uk

January 15, 2019

The database contains records for 1885 respondents. For each respondent 12 attributes are known: personality measurements which include 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. In addition, participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron) which was introduced to identify over-claimers. For each drug they selected either never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day.

The database contains 18 classification problems. Each of the independent label variables contains seven classes: 'Never Used', 'Used over a Decade Ago', 'Used in Last Decade', 'Used in Last Year', 'Used in Last Month', 'Used in Last Week', and 'Used in Last Day'.

Two versions of database is presented: original database with nominal input features and quantified database with numerical attributes.

Problems which can be solved:

- Seven class classifications for each drug separately.
- Problems can be transformed to binary classification by union of part of classes into one new class. For example, 'Never Used', 'Used over a Decade Ago' form class 'Non-user' and all other classes form class 'User'.
- The best binarization of classes for each attribute.
- Evaluation of risk to be drug consumer for specific drug.

The detailed description of the database is presented in:

1. [1] Fehrman, E., Egan, V., Gorban, A.N., Levesley, J., Mirkes, E.M., Muhammad, A.K. Personality Traits and Drug Consumption: The Story Told by Data, 2019, https://www.springer.com/gp/book/9783030104412
2. [2] Fehrman, E., Muhammad, A.K., Mirkes, E.M., Egan, V., Gorban, A.N., 2017. The Five Factor Model of personality and evaluation of drug consumption risk. In Data Science (pp. 231-242). Springer, Cham, https://doi.org/10.1007/978-3-319-55723-6_18
3. [3] Fehrman, E., Muhammad, A.K., Mirkes, E.M., Egan, V., Gorban, A.N., The Five Factor Model of personality and evaluation of drug consumption risk, arXiv preprint arXiv:1506.06297, 2015, https://arxiv.org/abs/1506.06297

Successful classifiers have been created for all drugs, thus providing the possibility of evaluating individuals for the risk of drug consumption. For most drugs sensitivity and specificity are greater than 75%.

Contents

Chapter 1

Data collection

The database was collected by Elaine Fehrman between March 2011 and March 2012. An online survey tool from Survey Gizmo was employed to gather data with maximum anonymity, this being particularly relevant to canvassing respondents' views, given the sensitive nature of drug use. All participants were required to declare themselves at least 18 years of age prior to informed consent being given.

The study recruited 2051 participants over an 12-month recruitment period. Of these persons, 166 did not respond correctly to a validity check built into the middle of the scale, so were presumed to being inattentive to the questions being asked. Nine of these persons were found to also have endorsed using a fictitious recreational drug, and which was included precisely to identify respondents who over-claim, as have other studies of this kind [4]. This led a useable sample of 1885 participants (male/female = 943/942).

The snowball sampling methodology recruited a primarily (93.5%) native English-speaking sample, with participants from the UK (1044; 55.4%), the USA (557; 29.5%), Canada (87; 4.6%), Australia (54; 2.9%), New Zealand (5; 0.3%) and Ireland (n = 20; 1.1%). A total of 118 (6.3%) came from a diversity of other countries, none of whom individually met 1% of the sample or did not declare the country of location. Further optimizing anonymity, persons reported their age band, rather than their exact age; 18-24 years (643; 34.1%), 25-34 years (481; 25.5%), 35-44 years (356; 18.9%), 45-54 years (294; 15.6%), 55-64 (93; 4.9%), and over 65 (18; 1%). This indicates that although the largest age cohort band were 18 to 24, some 40% of the cohort was 35 or above, which are a sample often missed in studies of this kind.

The sample recruited was highly educated, with just under two thirds (59.5%) educated to, at a minimum, degree or professional certificate level: 14.4% (271) reported holding a professional certificate or diploma, 25.5% (n = 481) an undergraduate degree, 15% (n = 284) a master's degree, and 4.7% (n = 89) a doctorate. Approximately 26.8% (n = 506) of the sample had received some college or university tuition although they did not hold any certificates; lastly, 257 (13.6%) had left school at the age of 18 or younger.

Participants were asked to indicate which racial category was broadly representative of their cultural background. An overwhelming majority (91.2%; 1720) reported being White, 1.8% (33) stated they were Black, and 1.4% (26) Asian. The remainder of the sample (5.6%; 106) described themselves as 'Other' or 'Mixed' categories. This small number of persons belonging to specific non-white ethnicities precludes any analyses involving racial categories.

1.1 Personality measurements

In order to assess personality traits of the sample, the NEO-FFI-R) questionnaire was employed [5]. The NEO-FFI-R is a highly reliable measure of basic personality domains; internal consistencies are 0.84 (N); 0.78 (E); 0.78 (O); 0.77 (A), and 0.75 (C) [6]. The scale is a 60-item inventory comprised of five personality domains or factors. The NEO-FFI-R is a shortened version of the Revised NEO-Personality Inventory (NEO-PI-R) [5]. The five factors are: N (Neuroticism), E (Extraversion), O (Openness), A (Agreeableness), and C (Conscientiousness) with 12 items per domain. These traits can be summarized as:

1. Neuroticism a long-term tendency to experience negative emotions such as nervousness, tension, anxiety and depression;

2. Extraversion manifested in outgoing, warm, active, assertive, talkative, cheerful, and in search of stimulation characteristics;

3. Openness a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests,

4. Agreeableness a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness;

5. Conscientiousness a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient.

All of these domains are hierarchically defined by specific facets [7]. [8] observe that the score Openness and Extraversion domains of the NEO-FFI instrument are less reliable than Neuroticism, Agreeableness, and Conscientiousness. Participants were asked to read the 60 NEO-FFI-R statements and indicate on a five-point Likert scale how much a given item applied to them (i.e. 0 = 'Strongly Disagree', 1 = 'Disagree', 2 = 'Neutral', 3 = 'Agree', to 4 = 'Strongly Agree').

We expected that drug usage is associated with high N, and low A and C. The darker dimension of personality can be described in terms of low A, whereas much of the anti-social behaviour in non-clinical persons appears underpinned by high N and low C [9]. The so-called 'negative urgency' is the tendency to act rashly when distressed, and characterized by high N, low C, and low A [10]. The negative urgency is partially proved below for users of most of the illegal drugs. In addition, our findings suggest that O is higher for drug users.

The second measure used was the Barratt Impulsiveness Scale (BIS-11) [11]. The BIS-11 is a 30-item self-report questionnaire, which measures the behavioural construct of impulsiveness, and comprises three subscales: motor impulsiveness, attentional impulsiveness, and non-planning. The 'motor' aspect reflects acting without thinking, the 'attentional' component poor concentration and thought intrusions, and the 'non-planning' a lack of consideration for consequences [12]. The scale's items are scored on a four-point Likert scale. This study modified the response range to make it compatible with previous related studies [13]. A score of five usually connotes the most impulsive response although some items are reverse-scored to prevent response bias. Items are aggregated, and the higher BIS-11 scores, the higher the impulsivity level [14]. The BIS-11 is regarded a reliable psychometric instrument with good test-retest reliability (Spearman's rho is equal to 0.83) and internal consistency (Cronbach's alpha is equal to 0.83; [11, 12].

The third measurement tool employed was the Impulsiveness Sensation-Seeking (ImpSS). Although the ImpSS combines the traits of impulsivity and sensation-seeking, it is regarded as a measure of a general sensation-seeking trait [15]. The scale consists of 19 statements in true-false format, comprising eight items measuring impulsivity (Imp), and 11 items gauging sensation-seeking (SS). The ImpSS is considered a valid and reliable measure of high risk behavioural correlates such as substance misuse [16].

1.2 Drug use

Participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine, and volatile substance abuse (VSA)) and one fictitious drug (Semeron) which was introduced to identify over-claimers.

It was recognised at the outset that drug use research regularly (and spuriously) dichotomises individuals as users or non-users, without due regard to their frequency or duration/desistance of drug use [17]. In this study, finer distinctions concerning the measurement of drug use have been deployed, due to the potential for the existence of qualitative differences amongst individuals with varying usage levels. In relation to each drug, respondents were asked to indicate on if they never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day. This format captured the breadth of a drug-using career, and the specific recency of use. The seven categories of drug users are depicted in Figure 1.1.

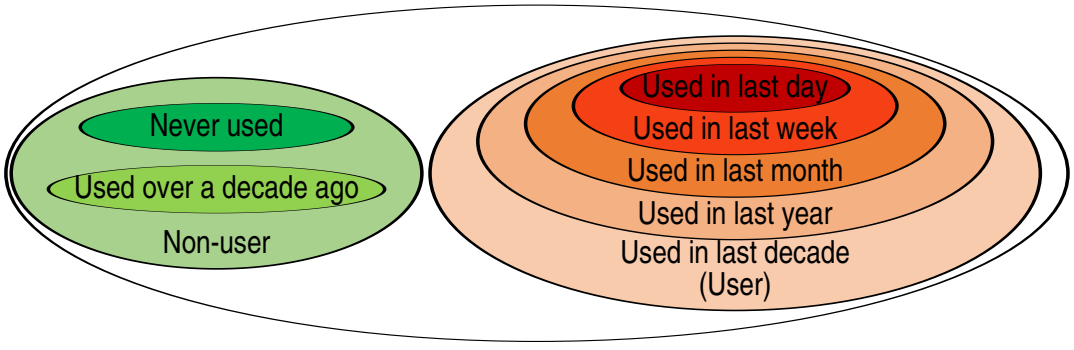


Figure 1.1: Categories of drug users

It can be seen that participants who had used a drug the previous day belong to the category ‘Used in last day’ and also to the categories ‘Used in last week’, ‘Used in last month’, ‘Used in last year’ and ‘Used in last decade’. There are two special categories (see Figure 1.1): ‘Never used’ and ‘Used over a decade ago’. These two categories were placed into the class of ‘Non-user’, and all other categories into the class ‘User’, as the simplest version of binary classification. Further in this study we analysed this binary classification.

The proportions of drug users differed for different drugs. The database sample comprised 1885 individuals without any missing data. Consumption of alcohol, caffeine, and chocolate was relatively common (over 96%). Consumption of cannabis and nicotine was also high (over 67%). Consumption of benzodiazepines, ecstasy, and legal highs was less, at 41%. Consumption of amphetamines, mushrooms and cocaine was approximately 36%. Consumption of ketamine and amyl nitrite is approximately 19%. Consumption of methadone is above 22% and LSD is less than 30%. Finally, crack, heroin, and VSA use is approximately 10%, 11%, 12%, respectively. These numbers characterise the group of respondents. It is worth to mention here that the sample is biased to the higher proportion of drug users and for the population consumption of the illegal drugs is expected to be significantly lower [18].

Chapter 2

Input feature transformation

There are many data mining methods to work with continuous data. It is necessary to quantify all categorical features to use these methods especially for features with big number of levels. Really, if we apply logistic regression for these data with categorical coefficients then we have to use dummy coding directly or indirectly. In this case we have n-1 coefficients for feature with n levels. It means that we fit logistic regression in the 250 dimensional space (age contains 6 levels, gender contains 2 levels, education contains 9 levels, country contains 7 levels, ethnicity contains 7 levels, Nscore contains 49 levels, Escore contains 42 levels, Oscore contains 35 levels, Ascore contains 41 levels, Cscore contains 41 levels, impulsive contains 10 levels, and SS contains 11 levels: 5+1+8+6+6+48+41+34+41+41+9+10=250). After quantification we can fit logistic regression model in 12 dimensional space. It means that feature quantification can be used as effective dimensionality reduction method.

2.1 Ordinal features quantification

One of the widely used techniques to analyse categorical data is the calculation of polychoric correlation [19, 20]. The matrix of polychoric coefficients further is used to calculate principal components, etc. The technique of polychoric correlation is based on suggestion that values of ordinal feature are the result of discretization of continuous random values with fixed thresholds. Furthermore, this latent continuous random value follows the normal distribution. Unfortunately, polychoric correlation techniques have two drawbacks: it defines the thresholds of discretization but not the values for each category and the defined thresholds are different for different pairs of attributes.

Let us have the ordinal feature O with categories $o_1, o_2, \dots o_k$, and with number of cases n_i of category o_i. The empirical estimation of probability of category o_i is $p_i = n_i/N$, where $N = \sum n_i$. The sample estimation of thresholds are evaluating as:

$$t_i = \Phi^{-1}\left(\sum_{j=1}^{i} p_j\right) \tag{2.1}$$

The simplest method of ordinal feature quantification is to use thresholds (2.1) and select the 'average' value in each interval. There are several variants of 'average' value. For this study we use the value with average probability: if thresholds t_{i-1} and t_i define the interval of category o_i, then average probability is

$$q_i = \Phi^{-1}\left(\sum_{j=1}^{i-1} p_j + \frac{p_i}{2}\right) \tag{2.2}$$

The polychoric coefficients, calculated on base of quantification (2.2), have less likelihood than polychoric coefficients calculated by using the maximum likelihood approach. The merit of this approach is the usage of the same thresholds for all pairs of attributes and explicit formula for calculation the categories' values.

2.2 Nominal feature quantification

We cannot use techniques described above to quantify nominal features such as gender, country of location and ethnicity because categories of these features are unordered. To quantify nominal features we implemented the technique of nonlinear CatPCA [21]. This procedure includes four steps:

1. Exclude nominal features from the set of input features and calculate the informative principal components [22–25] in space of retained input feature. To select informative components we use Kaiser's rule [26, 27].

2. Calculate the centroid of each category in projection on selected principal components.

3. Calculate the first principal component of centroids.

4. The numerical value for each component is the projection of its centroid on this component.

The process of nominal feature quantification for the feature 'Country' is depicted in Figure 2.1. Figure 2.1 shows that points corresponding to the UK category are located very far from any other points.

As an alternative variant of nominal feature quantification we use dummy coding [28] of nominal variables: 'country' is transformed into seven binary features with values 1 (if 'true') or 0 (if 'false'): UK, Canada, USA, Other (country), Australia, Republic of Ireland and New Zealand; Ethnicity is transformed into seven binary features: Mixed-White/Asian, White, Other (ethnicity), Mixed-White/Black, Asian, Black and Mixed-Black/Asian.

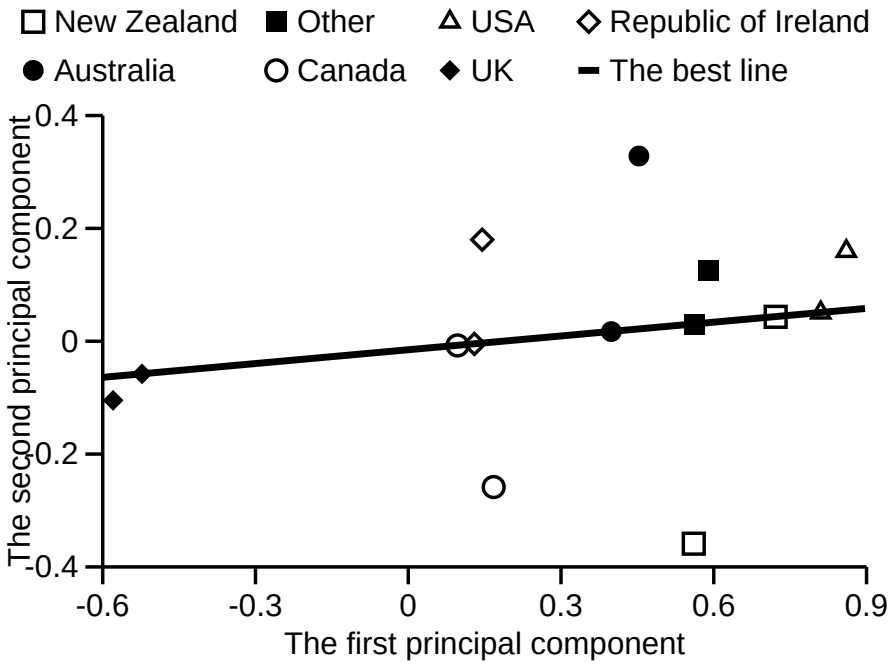


Figure 2.1: CatPCA quantification of 'Country' on the plane of the first two principal components

Chapter 3

Output features description

3.1 Alcohol

Alcohol is recency of alcohol consumption. It is output attribute with the following distribution of classes.

Table 3.1: Alcohol consumption

Value	Class	Cases	Fraction
CL0	Never Used	34	1.80%
CL1	Used over a Decade Ago	34	1.80%
CL2	Used in Last Decade	68	3.61%
CL3	Used in Last Year	198	10.50%
CL4	Used in Last Month	287	15.23%
CL5	Used in Last Week	759	40.27%
CL6	Used in Last Day	505	26.79%

3.2 Amphet

Amphet is recency of amphetamine consumption. It is output attribute with the following distribution of classes.

Table 3.2: Amphetamine consumption

Value	Class	Cases	Fraction
CL0	Never Used	976	51.78%
CL1	Used over a Decade Ago	230	12.20%
CL2	Used in Last Decade	243	12.89%
CL3	Used in Last Year	198	10.50%
CL4	Used in Last Month	75	3.98%
CL5	Used in Last Week	61	3.24%
CL6	Used in Last Day	102	5.41%

3.3 Amyl

Amyl is the recency of amyl nitrite consumption. It is output attribute with the following distribution of classes.

Table 3.3: Amyl nitrite consumption

Value	Class	Cases	Fraction
CL0	Never Used	1305	69.23%
CL1	Used over a Decade Ago	210	11.14%
CL2	Used in Last Decade	237	12.57%
CL3	Used in Last Year	92	4.88%
CL4	Used in Last Month	24	1.27%
CL5	Used in Last Week	14	0.74%
CL6	Used in Last Day	3	0.16%

3.4 Benzos

Benzos is recency of benzodiazepines consumption. It is output attribute with the following distribution of classes:

Table 3.4: Benzodiazepines consumption

Value	Class	Cases	Fraction
CL0	Never Used	1000	53.05%
CL1	Used over a Decade Ago	116	6.15%
CL2	Used in Last Decade	234	12.41%
CL3	Used in Last Year	236	12.52%
CL4	Used in Last Month	120	6.37%
CL5	Used in Last Week	84	4.46%
CL6	Used in Last Day	95	5.04%

3.5 Caff

Caff is recency of caffeine consumption. It is output attribute with the following distribution of classes.

Table 3.5: Caffeine consumption

Value	Class	Cases	Fraction
CL0	Never Used	27	1.43%
CL1	Used over a Decade Ago	10	0.53%
CL2	Used in Last Decade	24	1.27%
CL3	Used in Last Year	60	3.18%
CL4	Used in Last Month	106	5.62%
CL5	Used in Last Week	273	14.48%
CL6	Used in Last Day	1385	73.47%

3.6 Cannabis

Cannabis is recency of consumption. It is output attribute with the following distribution of classes.

Table 3.6: Cannabis consumption

Value	Class	Cases	Fraction
CL0	Never Used	413	21.91%
CL1	Used over a Decade Ago	207	10.98%
CL2	Used in Last Decade	266	14.11%
CL3	Used in Last Year	211	11.19%
CL4	Used in Last Month	140	7.43%
CL5	Used in Last Week	185	9.81%
CL6	Used in Last Day	463	24.56%

3.7 Choc

Choc is recency of chocolate consumption. It is output attribute with the following distribution of classes.

Table 3.7: Chocolate consumption

Value	Class	Cases	Fraction
CL0	Never Used	32	1.70%
CL1	Used over a Decade Ago	3	0.16%
CL2	Used in Last Decade	10	0.53%
CL3	Used in Last Year	54	2.86%
CL4	Used in Last Month	296	15.70%
CL5	Used in Last Week	683	36.23%
CL6	Used in Last Day	807	42.81%

3.8 Coke

Coke is recency of cocaine consumption. It is output attribute with the following distribution of classes:

Table 3.8: Cocaine consumption

Value	Class	Cases	Fraction
CL0	Never Used	1038	55.07%
CL1	Used over a Decade Ago	160	8.49%
CL2	Used in Last Decade	270	14.32%
CL3	Used in Last Year	258	13.69%
CL4	Used in Last Month	99	5.25%
CL5	Used in Last Week	41	2.18%
CL6	Used in Last Day	19	1.01%

3.9 Crack

Crack is recency of crack consumption. It is output attribute with the following distribution of classes.

Table 3.9: Crack consumption

Value	Class	Cases	Fraction
CL0	Never Used	1627	86.31%
CL1	Used over a Decade Ago	67	3.55%
CL2	Used in Last Decade	112	5.94%
CL3	Used in Last Year	59	3.13%
CL4	Used in Last Month	9	0.48%
CL5	Used in Last Week	9	0.48%
CL6	Used in Last Day	2	0.11%

3.10 Ecstasy

Ecstasy is recency of ecstasy consumption. It is output attribute with the following distribution of classes.

Table 3.10: Ecstasy consumption

Value	Class	Cases	Fraction
CL0	Never Used	1021	54.16%
CL1	Used over a Decade Ago	113	5.99%
CL2	Used in Last Decade	234	12.41%
CL3	Used in Last Year	277	14.69%
CL4	Used in Last Month	156	8.28%
CL5	Used in Last Week	63	3.34%
CL6	Used in Last Day	21	1.11%

3.11 Heroin

Heroin is recency of heroin consumption. It is output attribute with the following distribution of classes.

Table 3.11: Heroin consumption

Value	Class	Cases	Fraction
CL0	Never Used	1605	85.15%
CL1	Used over a Decade Ago	68	3.61%
CL2	Used in Last Decade	94	4.99%
CL3	Used in Last Year	65	3.45%
CL4	Used in Last Month	24	1.27%
CL5	Used in Last Week	16	0.85%
CL6	Used in Last Day	13	0.69%

3.12 Ketamine

Ketamine is recency of ketamine consumption. It is output attribute with the following distribution of classes:

Table 3.12: Ketamine consumption

Value	Class	Cases	Fraction
CL0	Never Used	1490	79.05%
CL1	Used over a Decade Ago	45	2.39%
CL2	Used in Last Decade	142	7.53%
CL3	Used in Last Year	129	6.84%
CL4	Used in Last Month	42	2.23%
CL5	Used in Last Week	33	1.75%
CL6	Used in Last Day	4	0.21%

3.13 Legalh

Legalh is the recency of “legal high” (no illegal) consumption. It is output attribute with the following distribution of classes

Table 3.13: Legal high consumption

Value	Class	Cases	Fraction
CL0	Never Used	1094	58.04%
CL1	Used over a Decade Ago	29	1.54%
CL2	Used in Last Decade	198	10.50%
CL3	Used in Last Year	323	17.14%
CL4	Used in Last Month	110	5.84%
CL5	Used in Last Week	64	3.40%
CL6	Used in Last Day	67	3.55%

3.14 LSD

LSD is recency of LSD consumption. It is output attribute with the following distribution of classes

Table 3.14: LSD consumption

Value	Class	Cases	Fraction
CL0	Never Used	1069	56.71%
CL1	Used over a Decade Ago	259	13.74%
CL2	Used in Last Decade	177	9.39%
CL3	Used in Last Year	214	11.35%
CL4	Used in Last Month	97	5.15%
CL5	Used in Last Week	56	2.97%
CL6	Used in Last Day	13	0.69%

3.15 Meth

Meth is recency of methadone consumption. It is output attribute with the following distribution of classes.

Table 3.15: Methadone consumption

Value	Class	Cases	Fraction
CL0	Never Used	1429	75.81%
CL1	Used over a Decade Ago	39	2.07%
CL2	Used in Last Decade	97	5.15%
CL3	Used in Last Year	149	7.90%
CL4	Used in Last Month	50	2.65%
CL5	Used in Last Week	48	2.55%
CL6	Used in Last Day	73	3.87%

3.16 Mushrooms

Mushrooms is recency of magic mushrooms consumption. It is output attribute with following distribution of classes:

Table 3.16: Magic mushrooms consumption

Value	Class	Cases	Fraction
CL0	Never Used	982	52.10%
CL1	Used over a Decade Ago	209	11.09%
CL2	Used in Last Decade	260	13.79%
CL3	Used in Last Year	275	14.59%
CL4	Used in Last Month	115	6.10%
CL5	Used in Last Week	40	2.12%
CL6	Used in Last Day	4	0.21%

3.17 Nicotine

Nicotine is recency of nicotine consumption. It is output attribute with the following distribution of classes.

Table 3.17: Nicotine consumption

Value	Class	Cases	Fraction
CL0	Never Used	428	22.71%
CL1	Used over a Decade Ago	193	10.24%
CL2	Used in Last Decade	204	10.82%
CL3	Used in Last Year	185	9.81%
CL4	Used in Last Month	108	5.73%
CL5	Used in Last Week	157	8.33%
CL6	Used in Last Day	610	32.36%

3.18 Semer

Semer is the recency of alleged consumption of a fictitious drug: Semeron. It is output attribute with the following distribution of classes.

Table 3.18: Semer consumption

Value	Class	Cases	Fraction
CL0	Never Used	1877	99.58%
CL1	Used over a Decade Ago	2	0.11%
CL2	Used in Last Decade	3	0.16%
CL3	Used in Last Year	2	0.11%
CL4	Used in Last Month	1	0.05%
CL5	Used in Last Week	0	0.00%
CL6	Used in Last Day	0	0.00%

3.19 VSA

VSA is recency of volatile substance consumption (e.g., solvents, petrol, etc). It is output attribute with the following distribution of classes:

Table 3.19: VSA consumption

Value	Class	Cases	Fraction
CL0	Never Used	1455	77.19%
CL1	Used over a Decade Ago	200	10.61%
CL2	Used in Last Decade	135	7.16%
CL3	Used in Last Year	61	3.24%
CL4	Used in Last Month	13	0.69%
CL5	Used in Last Week	14	0.74%
CL6	Used in Last Day	7	0.37%

Chapter 4

Original input features description (Original.csv)

Database does not contain missing values.

4.1 ID

ID is number of record in original database. It cannot be related to participant. It can be used for reference only.

4.2 Age

Age (Ordinal) is age of participant and has one of the following values:

Table 4.1: Age

Value	Meaning	Cases	Fraction
Ag1	18-24	643	34.11%
Ag2	25-34	481	25.52%
Ag3	35-44	356	18.89%
Ag4	45-54	294	15.60%
Ag5	55-64	93	4.93%
Ag6	65+	18	0.95%

4.3 Gender

Gender (Nominal) is the gender of participant: Value Meaning Cases Fraction

Table 4.2: Gender

Value	Meaning	Cases	Fraction
F	Female	942	49.97%
M	Male	943	50.03%

4.4 Education

Education (Ordinal) is level of education of participant and has one of the values:

Table 4.3: Education

Value	Meaning	Cases	Fraction
Ed1	Left school before 16 years	28	1.49%
Ed2	Left school at 16 years	99	5.25%
Ed3	Left school at 17 years	30	1.59%
Ed4	Left school at 18 years	100	5.31%
Ed5	Some college or university, no certificate or degree	506	26.84%
Ed6	Professional certificate/ diploma	270	14.32%
Ed7	University degree	480	25.46%
Ed8	Master degree	283	15.01%
Ed9	Doctorate degree	89	4.72%

4.5 Country

Country (Nominal) is country of current residence for the participant and has one of the following values:

Table 4.4: Country

Value	Meaning	Cases	Fraction
AU	Australia	54	2.86%
CA	Canada	87	4.62%
NZ	New Zealand	5	0.27%
OT	Other	118	6.26%
IE	Republic of Ireland	20	1.06%
UK	UK	1044	55.38%
US	USA	557	29.55%

4.6 Ethnicity

Ethnicity (Nominal) is the ethnicity of the participant and has one of these values:

Table 4.5: Ethnicity

Value	Meaning	Cases	Fraction
AS	Asian	26	1.38%
BL	Black	33	1.75%
BA	Mixed-Black/Asian	3	0.16%
WA	Mixed-White/Asian	20	1.06%
WB	Mixed-White/Black	20	1.06%
OT	Other	63	3.34%
WH	White	1720	91.25%

4.7 Nscore

Nscore (Interval) is the participants NEO-FFI-R Neuroticism raw score. Possible values can range from 12 to 60 (12 items rated from 1 to 5) and are presented in the table below. Variables up to measure 2.11 all follow this model.

Table 4.6: Neuroticism raw score

Value	Cases	Fraction	Value	Cases	Fraction	Value	Cases	Fraction
S12	1	0.05%	S29	60	3.18%	S46	67	3.55%
S13	1	0.05%	S30	61	3.24%	S47	27	1.43%
S14	7	0.37%	S31	87	4.62%	S48	49	2.60%
S15	4	0.21%	S32	78	4.14%	S49	40	2.12%
S16	3	0.16%	S33	68	3.61%	S50	24	1.27%
S17	4	0.21%	S34	76	4.03%	S51	27	1.43%
S18	10	0.53%	S35	69	3.66%	S52	17	0.90%
S19	16	0.85%	S36	73	3.87%	S53	20	1.06%
S20	24	1.27%	S37	67	3.55%	S54	15	0.80%
S21	31	1.64%	S38	63	3.34%	S55	11	0.58%
S22	26	1.38%	S39	66	3.50%	S56	10	0.53%
S23	29	1.54%	S40	80	4.24%	S57	6	0.32%
S24	35	1.86%	S41	61	3.24%	S58	3	0.16%
S25	56	2.97%	S42	77	4.08%	S59	5	0.27%
S26	57	3.02%	S43	49	2.60%	S60	2	0.11%
S27	65	3.45%	S44	51	2.71%			
S28	70	3.71%	S45	37	1.96%			

4.8 Escore

Escore (Interval) is NEO-FFI-R Extraversion. Possible values are presented in the table below.

Table 4.7: Extraversion raw score

Value	Cases	Fraction	Value	Cases	Fraction	Value	Cases	Fraction
S16	2	0.11%	S31	55	2.92%	S45	91	4.83%
S18	1	0.05%	S32	52	2.76%	S46	69	3.66%
S19	6	0.32%	S33	77	4.08%	S47	64	3.40%
S20	3	0.16%	S34	68	3.61%	S48	62	3.29%
S21	3	0.16%	S35	58	3.08%	S49	37	1.96%
S22	8	0.42%	S36	89	4.72%	S50	25	1.33%
S23	5	0.27%	S37	90	4.77%	S51	34	1.80%
S24	9	0.48%	S38	106	5.62%	S52	21	1.11%
S25	4	0.21%	S39	107	5.68%	S53	15	0.80%
S26	21	1.11%	S40	130	6.90%	S54	10	0.53%
S27	23	1.22%	S41	116	6.15%	S55	9	0.48%
S28	23	1.22%	S42	109	5.78%	S56	2	0.11%
S29	32	1.70%	S43	105	5.57%	S58	1	0.05%
S30	38	2.02%	S44	103	5.46%	S59	2	0.11%

4.9 Oscore

Oscore (Interval) is NEO-FFI-R Openness to experience. Possible values are presented in the table below.

Table 4.8: Openness to experience raw score

Value	Cases	Fraction	Value	Cases	Fraction	Value	Cases	Fraction
S24	2	0.11%	S38	64	3.40%	S50	83	4.40%
S26	4	0.21%	S39	60	3.18%	S51	87	4.62%
S28	4	0.21%	S40	68	3.61%	S52	87	4.62%
S29	11	0.58%	S41	76	4.03%	S53	81	4.30%
S30	9	0.48%	S42	87	4.62%	S54	57	3.02%
S31	9	0.48%	S43	86	4.56%	S55	63	3.34%
S32	13	0.69%	S44	101	5.36%	S56	38	2.02%
S33	23	1.22%	S45	103	5.46%	S57	34	1.80%
S34	25	1.33%	S46	134	7.11%	S58	19	1.01%
S35	26	1.38%	S47	107	5.68%	S59	13	0.69%
S36	39	2.07%	S48	116	6.15%	S60	7	0.37%
S37	51	2.71%	S49	98	5.20%			

4.10 Ascore

Ascore (Interval) is NEO-FFI-R Agreeableness. Possible values are presented in the table below.

Table 4.9: Agreeableness raw score

Value	Cases	Fraction	Value	Cases	Fraction	Value	Cases	Fraction
S12	1	0.05%	S34	42	2.23%	S48	104	5.52%
S16	1	0.05%	S35	45	2.39%	S49	85	4.51%
S18	1	0.05%	S36	62	3.29%	S50	68	3.61%
S23	1	0.05%	S37	83	4.40%	S51	58	3.08%
S24	2	0.11%	S38	82	4.35%	S52	39	2.07%
S25	1	0.05%	S39	102	5.41%	S53	36	1.91%
S26	7	0.37%	S40	98	5.20%	S54	36	1.91%
S27	7	0.37%	S41	114	6.05%	S55	16	0.85%
S28	8	0.42%	S42	101	5.36%	S56	14	0.74%
S29	13	0.69%	S43	105	5.57%	S57	8	0.42%
S30	18	0.95%	S44	118	6.26%	S58	7	0.37%
S31	24	1.27%	S45	112	5.94%	S59	1	0.05%
S32	30	1.59%	S46	100	5.31%	S60	1	0.05%
S33	34	1.80%	S47	100	5.31%			

4.11 Cscore

Cscore (Interval) is NEO-FFI-R Conscientiousness. Possible values are presented in the table below:

Table 4.10: Conscientiousness raw score

Value	Cases	Fraction	Value	Cases	Fraction	Value	Cases	Fraction
S17	1	0.05%	S32	39	2.07%	S46	113	5.99%
S19	1	0.05%	S33	49	2.60%	S47	95	5.04%
S20	3	0.16%	S34	55	2.92%	S48	95	5.04%
S21	2	0.11%	S35	55	2.92%	S49	76	4.03%
S22	5	0.27%	S36	69	3.66%	S50	47	2.49%
S23	5	0.27%	S37	81	4.30%	S51	43	2.28%
S24	6	0.32%	S38	77	4.08%	S52	34	1.80%
S25	9	0.48%	S39	87	4.62%	S53	28	1.49%
S26	13	0.69%	S40	97	5.15%	S54	27	1.43%
S27	13	0.69%	S41	99	5.25%	S55	13	0.69%
S28	25	1.33%	S42	105	5.57%	S56	8	0.42%
S29	24	1.27%	S43	90	4.77%	S57	3	0.16%
S30	29	1.54%	S44	111	5.89%	S59	1	0.05%
S31	41	2.18%	S45	111	5.89%			

4.12 SS

SS (Ordinal) is sensation-seeking measured by the ImpSS measure. Possible values are presented in the table below:

Table 4.11: Conscientiousness (left) and impulsiveness (right) raw score

Value	Cases	Fraction
SS00	71	3.77%
SS01	87	4.62%
SS02	132	7.00%
SS03	169	8.97%
SS04	211	11.19%
SS05	223	11.83%
SS06	219	11.62%
SS07	249	13.21%
SS08	211	11.19%
SS09	210	11.14%
SS10	103	5.46%

Value	Cases	Fraction
I0	20	1.06%
I1	276	14.64%
I2	307	16.29%
I3	355	18.83%
I4	257	13.63%
I5	216	11.46%
I6	195	10.34%
I7	148	7.85%
I8	104	5.52%
I9	7	0.37%

4.13 Impulsivity

Impulsivity (Ordinal) is impulsiveness as measured by BIS-11. Possible values are presented in the table above.

Chapter 5

Quantified input features description (Quantified.csv)

Database does not contain missing values. Tables with descriptive statistics present minimal (Min), maximal (Max), and mean (Mean) values and standard deviation (STD).

5.1 ID

ID is number of record in original database. It cannot be related to participant. It can be used for reference only.

5.2 Age

Age is age of participant. This feature is quantified from ordinal original.

Table 5.1: Age

Value	Meaning	Cases	Fraction
-0.95197	18-24	643	34.11%
-0.07854	25-34	481	25.52%
0.49788	35-44	356	18.89%
1.09449	45-54	294	15.60%
1.82213	55-64	93	4.93%
2.59171	65+	18	0.95%

Descriptive statistics

Min	Max	Mean	STD
-0.95197	2.59171	0.03461	0.87813

5.3 Gender

Gender is the gender of participant. This feature is quantified from nominal original.

Table 5.2: Gender

Value	Meaning	Cases	Fraction
0.48246	Female	942	49.97%
-0.48246	Male	943	50.03%

Descriptive statistics

Min	Max	Mean	STD
-0.48246	0.48246	-0.00026	0.48246

5.4 Education

Education is level of education of participant. This feature is quantified from ordinal original.

Table 5.3: Education

Value	Meaning	Cases	Fraction
-2.43591	Left school before 16 years	28	1.49%
-1.73790	Left school at 16 years	99	5.25%
-1.43719	Left school at 17 years	30	1.59%
-1.22751	Left school at 18 years	100	5.31%
-0.61113	Some college or university, no certificate or degree	506	26.84%
-0.05921	Professional certificate/ diploma	270	14.32%
0.45468	University degree	480	25.46%
1.16365	Masters degree	283	15.01%
1.98437	Doctorate degree	89	4.72%

Descriptive statistics

Min	Max	Mean	STD
-2.43591	1.98437	-0.00379	0.95004

5.5 Country

Country is country of current residence for the participant. This feature is quantified from nominal original.

Table 5.4: Age

Value	Meaning	Cases	Fraction
-0.09765	Australia	54	2.86%
0.24923	Canada	87	4.62%
-0.46841	New Zealand	5	0.27%
-0.28519	Other	118	6.26%
0.21128	Republic of Ireland	20	1.06%
0.96082	UK	1044	55.38%
-0.57009	USA	557	29.55%

Descriptive statistics

Min	Max	Mean	STD
-0.57009	0.96082	0.35554	0.70015

5.6 Ethnicity

Ethnicity is the ethnicity of the participant. This feature is quantified from nominal original.

Table 5.5: Age

Value	Meaning	Cases	Fraction
-0.50212	Asian	26	1.38%
-1.10702	Black	33	1.75%
1.90725	Mixed-Black/Asian	3	0.16%
0.12600	Mixed-White/Asian	20	1.06%
-0.22166	Mixed-White/Black	20	1.06%
0.11440	Other	63	3.34%
-0.31685	White	1720	91.25%

Descriptive statistics

Min	Max	Mean	STD
-1.10702	1.90725	-0.30958	0.16618

5.7 Nscore

Nscore is NEO-FFI-R Neuroticism. This feature is quantified from interval original.

Table 5.6: Neuroticism

Score	Cases	Value	Score	Cases	Value	Score	Cases	Value
12	1	-3.46436	29	60	-0.67825	46	67	1.02119
13	1	-3.15735	30	61	-0.58016	47	27	1.13281
14	7	-2.75696	31	87	-0.46725	48	49	1.23461
15	4	-2.52197	32	78	-0.34799	49	40	1.37297
16	3	-2.42317	33	68	-0.24649	50	24	1.49158
17	4	-2.34360	34	76	-0.14882	51	27	1.60383
18	10	-2.21844	35	69	-0.05188	52	17	1.72012
19	16	-2.05048	36	73	0.04257	53	20	1.83990
20	24	-1.86962	37	67	0.13606	54	15	1.98437
21	31	-1.69163	38	63	0.22393	55	11	2.12700
22	26	-1.55078	39	66	0.31287	56	10	2.28554
23	29	-1.43907	40	80	0.41667	57	6	2.46262
24	35	-1.32828	41	61	0.52135	58	3	2.61139
25	56	-1.19430	42	77	0.62967	59	5	2.82196
26	57	-1.05308	43	49	0.73545	60	2	3.27393
27	65	-0.92104	44	51	0.82562			
28	70	-0.79151	45	37	0.91093			

Descriptive statistics

Min	Max	Mean	STD
-3.46436	3.27393	0.00004	0.99808

5.8 Escore

Escore is NEO-FFI-R Extraversion. This feature is quantified from interval original.

Table 5.7: Extraversion

Score	Cases	Value	Score	Cases	Value	Score	Cases	Value
16	2	-3.27393	31	55	-1.23177	45	91	0.80523
18	1	-3.00537	32	52	-1.09207	46	69	0.96248
19	6	-2.72827	33	77	-0.94779	47	64	1.11406
20	3	-2.53830	34	68	-0.80615	48	62	1.28610
21	3	-2.44904	35	58	-0.69509	49	37	1.45421
22	8	-2.32338	36	89	-0.57545	50	25	1.58487
23	5	-2.21069	37	90	-0.43999	51	34	1.74091
24	9	-2.11437	38	106	-0.30033	52	21	1.93886
25	4	-2.03972	39	107	-0.15487	53	15	2.12700
26	21	-1.92173	40	130	0.00332	54	10	2.32338
27	23	-1.76250	41	116	0.16767	55	9	2.57309
28	23	-1.63340	42	109	0.32197	56	2	2.85950
29	32	-1.50796	43	105	0.47617	58	1	3.00537
30	38	-1.37639	44	103	0.63779	59	2	3.27393

Descriptive statistics

Min	Max	Mean	STD
-3.27393	3.27393	-0.00016	0.99745

5.9 Oscore

Oscore is NEO-FFI-R Openness to experience. This feature is quantified from interval original.

Table 5.8: Openness to experience

Score	Cases	Value	Score	Cases	Value	Score	Cases	Value
24	2	-3.27393	38	64	-1.11902	50	83	0.58331
26	4	-2.85950	39	60	-0.97631	51	87	0.72330
28	4	-2.63199	40	68	-0.84732	52	87	0.88309
29	11	-2.39883	41	76	-0.71727	53	81	1.06238
30	9	-2.21069	42	87	-0.58331	54	57	1.24033
31	9	-2.09015	43	86	-0.45174	55	63	1.43533
32	13	-1.97495	44	101	-0.31776	56	38	1.65653
33	23	-1.82919	45	103	-0.17779	57	34	1.88511
34	25	-1.68062	46	134	-0.01928	58	19	2.15324
35	26	-1.55521	47	107	0.14143	59	13	2.44904
36	39	-1.42424	48	116	0.29338	60	7	2.90161
37	51	-1.27553	49	98	0.44585			

Descriptive statistics

Min	Max	Mean	STD
-3.27393	2.90161	-0.00053	0.99623

5.10 Ascore

Ascore is NEO-FFI-R Agreeableness. This feature is quantified from interval original.

Table 5.9: Agreeableness

Score	Cases	Value	Score	Cases	Value	Score	Cases	Value
12	1	-3.46436	34	42	-1.34289	48	104	0.76096
16	1	-3.15735	35	45	-1.21213	49	85	0.94156
18	1	-3.00537	36	62	-1.07533	50	68	1.11406
23	1	-2.90161	37	83	-0.91699	51	58	1.2861
24	2	-2.78793	38	82	-0.76096	52	39	1.45039
25	1	-2.70172	39	102	-0.60633	53	36	1.61108
26	7	-2.5383	40	98	-0.45321	54	36	1.81866
27	7	-2.35413	41	114	-0.30172	55	16	2.03972
28	8	-2.21844	42	101	-0.15487	56	14	2.23427
29	13	-2.07848	43	105	-0.01729	57	8	2.46262
30	18	-1.92595	44	118	0.13136	58	7	2.75696
31	24	-1.772	45	112	0.28783	59	1	3.15735
32	30	-1.6209	46	100	0.43852	60	1	3.46436
33	34	-1.47955	47	100	0.59042			

Descriptive statistics

Min	Max	Mean	STD
-3.46436	3.46436	-0.00024	0.99744

5.11 Cscore

Cscore is NEO-FFI-R Conscientiousness. This feature is quantified from interval original.

Table 5.10: Conscientiousness

Score	Cases	Value	Score	Cases	Value	Score	Cases	Value
17	1	-3.46436	32	39	-1.25773	46	113	0.58489
19	1	-3.15735	33	49	-1.13788	47	95	0.7583
20	3	-2.90161	34	55	-1.0145	48	95	0.93949
21	2	-2.72827	35	55	-0.89891	49	76	1.13407
22	5	-2.57309	36	69	-0.78155	50	47	1.30612
23	5	-2.42317	37	81	-0.65253	51	43	1.46191
24	6	-2.30408	38	77	-0.52745	52	34	1.63088
25	9	-2.18109	39	87	-0.40581	53	28	1.81175
26	13	-2.04506	40	97	-0.27607	54	27	2.04506
27	13	-1.92173	41	99	-0.14277	55	13	2.33337
28	25	-1.78169	42	105	-0.00665	56	8	2.63199
29	24	-1.64101	43	90	0.12331	57	3	3.00537
30	29	-1.5184	44	111	0.25953	59	1	3.46436
31	41	-1.38502	45	111	0.41594			

Descriptive statistics

Min	Max	Mean	STD
-3.46436	3.46436	-0.00039	0.99752

5.12 SS

SS is sensation seeing measured by ImpSS. This feature is quantified from ordinal original.

Table 5.11: Sensation seeing

Value	Score	Cases	Fraction
-2.07848	0	71	3.77%
-1.54858	1	87	4.62%
-1.18084	2	132	7.00%
-0.84637	3	169	8.97%
-0.52593	4	211	11.19%
-0.21575	5	223	11.83%
0.07987	6	219	11.62%
0.40148	7	249	13.21%
0.76540	8	211	11.19%
1.22470	9	210	11.14%
1.92173	10	103	5.46%

Descriptive statistics

Min	Max	Mean	STD
-2.07848	1.92173	-0.00329	0.96370

5.13 Impulsivity

Impulsivity is impulsiveness measured by BIS-11. This feature is quantified from ordinal original.

Table 5.12: Impulsiveness

Value	Score	Cases	Fraction
-2.55524	0	20	1.06%
-1.37983	1	276	14.64%
-0.71126	2	307	16.29%
-0.21712	3	355	18.83%
0.19268	4	257	13.63%
0.52975	5	216	11.46%
0.88113	6	195	10.34%
1.29221	7	148	7.85%
1.86203	8	104	5.52%
2.90161	9	7	0.37%

Descriptive statistics

Min	Max	Mean	STD
-2.55524	2.90161	0.00721	0.95446

Bibliography

1. Fehrman, E., Egan, V., Gorban, A.N., Levesley, J., Mirkes, E.M., Muhammad, A.K. Personality Traits and Drug Consumption: The Story Told by Data, 2019, https://www.springer.com/gp/book/9783030104412

2. Fehrman, E., Muhammad, A.K., Mirkes, E.M., Egan, V., Gorban, A.N., 2017. The Five Factor Model of personality and evaluation of drug consumption risk. In Data Science (pp. 231-242). Springer, Cham, https://doi.org/10.1007/978-3-319-55723-6_18

3. Fehrman, E., Muhammad, A.K., Mirkes, E.M., Egan, V., Gorban, A.N., The Five Factor Model of personality and evaluation of drug consumption risk, arXiv preprint arXiv:1506.06297, 2015, https://arxiv.org/abs/1506.06297

4. Hoare, J., Moon, D. (eds.) Drug misuse declared: findings from the 2009/10 British Crime Survey. Home Office Statistical Bulletin 13/10 (2010). Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/116321/hosb1310.pdf. Accessed 27 Dec 2017

5. Costa, P.T., MacCrae, R.R.: Revised NEO-personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual. Psychological Assessment Resources, Odessa, FL (1992)

6. Egan, V.: Individual differences and antisocial behaviour. In: Furnham, A., Stumm, S., Petredies, K., (eds.): The Blackwell-Wiley Handbook of Individual Differences, pp. 512–537. Blackwell-Wiley, Oxford (2011). https://doi.org /10.1002/9781444343120.ch19

7. McCrae, R.R., Costa, P.T.: The NEO personality inventory: using the five-factor model in counseling. J. Couns. Dev. 69(4), 367–372 (1991). https://doi.org/10.1002/j.1556-6676.1991.tb01524.x

8. Egan, V., Deary, I., Austin, E.: The NEO-FFI: Emerging British norms and an item-level analysis suggest N, A and C are more reliable than O and E. Personal. Individ. Differ. 29(5), 907–920 (2000). https://doi.org/10.1016/s0191-8869(99)00242-1

9. Jakobwitz, S., Egan, V.: The dark triad and normal personality traits. Personal. Individ. Differ. 40(2), 331–339 (2006). https://doi.org/10.1016/j.paid.2005.07.006

10. Settles, R.E., Fischer, S., Cyders, M.A., Combs, J.L., Gunn, R.L., Smith G.T.: Negative urgency: a personality predictor of externalizing behavior characterized by neuroticism, low conscientiousness, and disagreeableness. J. Abnorm. Psychol. 121(1), 160–172 (2012). https://doi.org/10.1037/a0024948

11. Stanford, M.S., Mathias, C.W., Dougherty, D.M., Lake, S.L., Anderson, N.E., Patton, J.H.: Fifty years of the Barratt Impulsiveness Scale: An update and review. Personal. Individ. Differ. 47(5), 385–395 (2009). https://doi.org/10.1016/j.paid.2009.04.008

12. Snowden, R.J., Gray, N.S.: Impulsivity and psychopathy: Associations between the barrett impulsivity scale and the psychopathy checklist revised. Psychiatry Research. 87(3), 414–417 (2011). https://doi.org/10.1016/j.psychres .2011.02.003

13. García-Montes, J.M., Zaldívar-Basurto, F., López-Ríos, F., Molina-Moreno, A.: The role of personality variables in drug abuse in a Spanish university population. Int. J. Ment. Health Addict. 7(3), 475–487 (2009). https://doi.org/10.1007/ s11469-007-9144-y

14. Fossati, P., Ergis, A.M., Allilaire, J.F.: Problem-solving abilities in unipolar depressed patients: comparison of performance on the modified version of the Wisconsin and the California sorting tests. Psychiatry Res. 104(2), 145–156 (2001). https://doi.org/10.1016/s0165-1781(01)00307-9

15. Zuckerman, M.: Behavioral expressions and biosocial bases of sensation seeking. Cambridge University Press, New York (1994).

16. McDaniel, S.R., Mahan, J.E.: An examination of the Impss scale as a valid and reliable alternative to the SSS-V in optimum stimulation level research. Personal. Individ. Differ. 44(7), 1528–1538 (2008). https://doi.org/10.1016/j.paid.2008.01.009

17. Ragan, D.T., Beaver, K.M.: Chronic offenders: a life-course analysis of marijuana Users. Youth & Soc. 42(2), 174–198 (2010). https://doi.org/10.1177/0044 118x09351788

18. Home Office UK: Drug misuse: findings from the 2013 to 2014 Crime Survey for England and Wales, 2nd edn. (2014). Available: https://www.gov.uk/govern ment/statistics/drug-misuse-findings-from-the-2013-to-2014-csew. Accessed 27 Dec 2017

19. Lee, S.Y., Poon, W.Y., Bentler, P.M.: A two-stage estimation of structural equation models with continuous and polytomous variables. Br. J. Math. Stat. Psychol. 48(2), 339–358 (1995). https://doi.org/10.1111/j.2044-8317.1995.tb01067.x

20. Martinson, E.O., Hamdan, M.A.: Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables. J. Stat. Comput. Simul. 1(1), 45–54 (1971). https://doi.org/10.1080/009496572088 10003

21. Linting, M., van der Kooij, A.: Nonlinear principal components analysis with CATPCA: A tutorial. J. Pers. Assess. 94(1), 12–25 (2012). https://doi.org/10.1080/00223891.2011.627965

22. Pearson, K.: On lines and planes of closest fit to system of points in space. Philos. Mag. Ser. 6 2(11), 559–572 (1901). https://doi.org/10.1080/14786440109462720

23. Gorban, A.N., Zinovyev, A.Y.: Principal manifolds and graphs in practice: from molecular biology to dynamical systems. Int. j. neural syst. 20(03), 219–232 (2010). https://doi.org/10.1142/s0129065710002383

24. Gorban, A.N., Kégl, B., Wunsch, D.C., Zinovyev, A.Y. (eds.): Principal Manifolds for Data Visualisation and Dimension Reduction. LNCSE, vol. 58, Springer, Berlin-Heidelberg-New York (2008). https://doi.org/10.1007/978-3-540-73750-6

25. Gorban, A.N., Zinovyev, A.Y.: Principal graphs and manifolds. In: Olivas, E.S., Guerrero, J.D.M., Sober, M.M., Benedito, J.R.M., López, A.J.S. (eds.): Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. pp. 28–59. IGI Global, Hershey - New York (2009). https://doi.org/10.4018/978-1-60566-766-9.ch002

26. Guttman, L.: Some necessary conditions for common-factor analysis. Psychometrika 19(2), 149–161 (1954). https://doi.org/10.1007/bf02289162

27. Kaiser, H.F.: The application of electronic computers to factor analysis. Educ. Psychol. Meas. 20(1), 141–151 (1960). https://doi.org/10.1177/001316446002000116

28. Gujarati, D.N.: Basic econometrics, 4th edn. McGraw-Hill, New York (2004)

29. Terracciano, A., Lóckenhoff, C.E., Crum, R.M., Bienvenu, O.J., Costa, P.T.: Five-factor model personality profiles of drug users. BMC Psychiatr. 8(1), 22 (2008). https://doi.org/10.1186/1471-244x-8-22