Model selection, union and assembling in practical data analysis: Methods and case study

Thesis submitted to the University of Leicester

for the degree of Doctor of Philosophy

By

Awaz K. Muhammad

Department of Mathematics

University of Leicester

United Kingdom

June 22, 2018

Abstract

The main problem in KDD (Knowledge Discovery and Data Mining) is always two-fold: we have to discover knowledge in real data and we need to develop methods for KDD. This thesis is also two-fold.

First, I participated in the support and maintenance of the project 'Personality traits and drug consumption'. The real data from almost 2000 respondents have been analysed. My role was in data analysis and risk assessment. The central problem is in the search and validation of psychological predictors of consumption of different drugs. Eight data mining algorithms were used for user/non-user classification: decision trees, random forests, k-nearest neighbours, linear discriminant analysis, Gaussian mixtures, probability density function estimation by radial basis functions, logistic regression, and naïve Bayes. Correlation analysis based on the Pearson's correlation coefficient and on relative information gain revealed the existence of groups of drugs with strongly correlated consumption. Three correlation pleiades were identified.Classifiers with sensitivity and specificity being greater than 70% for almost all classification tasks were obtained.

Secondly, several new methods and approaches to feature selection were proposed and tested on the drug consumption database and on several other publicly available databases. These methods include 'double Kaiser selection' for selection of the main factors (principal components) and main attributes. Consideration of each attribute as a distribution on factors allowed us to apply any Kaiser rule for feature selection as well. We developed a methodology for creation and utilisation controllable multicollinearity. Multicollinearity can be useful because it allows to correct mistakes in data and to evaluate missed data. It is undesirable because many statistical tasks become ill-conditional. Alternative attribute sets approach (AASA) can determine several sets of relevant attributes that can be used to solve original problems separately. We tested AASA on several classification problems. We demonstrated that this methodology could be more accurate than the best traditional feature selection methods.

Disclaimer

This PhD is purely about data analysis. The data received was from experimental psychologists. The interpretation of the results of data analysis with regards to human behaviour is the domain of psychologists and is not the subject of this thesis.

Acknowledgements

It is my great pleasure to take this opportunity to acknowledge all the support I have received during my PhD study.

First of all, I would like to thank my supervisor and scientific father ***Professor Alexander Gorban***, for his generous support, encouragement, and invaluable advice at all the stages of this study. I am really indebted to him for admitting me into the department and under his supervision. His wealth of experience and knowledge I would never have had received anywhere else.

I particularly thank to my second supervisor ***Dr. Evgeny M. Mirkes*** for his kindly support, and invaluable guidance during my study. He contributed to my understanding of the topic in several ways.

I wish to thank ***Elaine Fehrman*** and her supervisor, ***Professor Vincent Egan*** for giving me permission to use the 'Drug consumption' database. I am very grateful to have ***Dr. Andrei Zinovyev*** for his expertise in the ViDaExpert's software.

Thanks to all staff in the College House, University of Leicester, for their help.

Also my appreciation goes to my lovely father and mother. Special thanks to my parents in law for their support. To my most wonderful husband Dzhwar, words are insufficient to express my sincere appreciation.

I wish to express my gratitude to my sponsor, the Ministry of Higher Education, Kurdistan-Iraq and Salahaddin University.

Finally, I wish to express my personal gratitude to my family and friends, who generously provided me with advice and support from the initial stages to the end of my research journey.

Contents

1 Introduction **1**

1.1 Backgrounds and analysis of dataset 1

1.2 Definitions of drugs and drug usage 3

1.3 Personality traits 5

1.4 The problem of relations between personality traits and drug consumption 7

1.5 New dataset 'drug consumption' 10

1.6 First results in brief 11

1.7 The reviews of feature selection problem 15

1.8 A methodology for selection of alternative sets of attributes 17

1.9 Model construction for Alternative Attributes Sets (AAS) 18

1.10 AASA applications to the datasets 19

1.11 Thesis Outline 20

2 Data **24**

2.1 Database 24

2.2 Personality measurements 27

2.3 Drugs and categories of users 31

2.4 Data sets for AASA testing 35

3 Materials and Methods of Data Analysis **36**

3.1 T–scores 36

3.2 Input feature transformation 37

3.2.1 Principal Component Analysis – PCA 38

3.2.2 Ordinal feature quantification 40

3.2.3 Nominal feature quantification 41

3.2.4 Input feature ranking 42

3.3 Methods of classification and risk evaluation 43

3.3.1 *k* Nearest Neighbours (*k*NN) 44

3.3.2 Decision Tree (DT) 49

3.3.3 Linear Discriminant Analysis (LDA) 52

3.3.4 Gaussian Mixture (GM) 52

3.3.5 Probability Density Function Estimation (PDFE) 53

3.3.6 Logistic Regression (LR) 54

3.3.7 Naïve Bayes (NB) 55

3.3.8 Random Forest (RF) 56

3.4 Criterion for selecting the best method 58

3.5 Visualisation of the non-linear principal components screen: Elastic maps 59

3.6 Alternative Attributes Set (AAS) 61

4 Results of data analysis **70**

4.1 Descriptive statistics and psychological profile of illicit drug users . 70

4.2 Distribution of number of drugs used 76

4.3 Sample mean and population norm 79

4.4 Deviation of the groups of drug users from the sample mean 81
4.5 Significant differences between groups of drug users and non-users 85
4.6 Correlation between usage of different drugs 91
4.7 Input feature ranking 96
4.8 Selection of the best classifiers for the decade-based classification problem 99
4.9 The best class binarization 104
4.10 Correlation pleiades of drugs 109
4.11 Overoptimism problem 116
4.12 User/non-user classification by linear discriminant for ecstasy and heroin 119
4.13 Separation of heroin users from ecstasy users 124
4.14 A tree of linear discriminants 130
4.15 Visualisation on non-linear PCA screen 132
4.16 Risk maps 134
4.17 Discussion 141

5 Alternative Attributes Sets Approach (AASA) 149

5.1 Introduction 149
5.2 Standard feature selection approaches 151
5.3 Time consumption of three classical algorithms 152
5.4 Alternative Attributes Set (AAS) 153
5.5 AASA with usage of several classifiers 156
5.6 AASA for drug consumption dataset 156
5.6.1 AASA for heroin usage by using LR and ES, FFS, and BFS . 157

5.6.2 AASA for cannabis and ecstasy consumption by usage LR and ES, FFS, and BFS 165

5.6.3 AASA for ecstasy consumption by usage of several classifiers and ES, FFS, and BFS 168

5.6.4 AASA for cannabis consumption by usage of several classifiers and ES, FFS, and BFS 175

5.7 AASA for USA president elections dataset 177

5.7.1 AASA for USA president elections by usage LR and ES, FFS, and BFS 177

5.7.2 AASA for USA president elections by usage of several classifiers and ES, FFS, and BFS 178

5.8 AASA for Breast cancer dataset 180

5.8.1 AASA for breast cancer by usage LR and FFS and BFS . . . 181

5.8.2 AASA for breast cancer by usage of several classifiers and FFS and BFS 184

5.9 Discussion 184

6 Conclusion and outlook **187**

Appendix A Flowcharts of classification methods **194**

Appendix B Flowcharts of AASA for feature selection **211**

Appendix C Main tables **229**

C.1 Psychological profiles of drug users and non-users 229

C.2 Correlation between consumption of different drugs 262

C.3 Linear discriminants for user/non-user separation 265

Appendix D Minimal feature sets **283**

D.1 Minimal feature sets for data set I (drug consumption) for heroin, ecstasy, and cannabis consumption. 283

D.2 Minimal feature sets for data set II (president elections of USA) . . 290

List of Tables

2.1 Country of residence and ethnicity composition 26

2.2 The age bands and education level 27

2.3 Pearson's correlation coefficients (PCC) between NEO-FFI trait scores for a large British sample, $n = 1025$ [41]; the p-value is the probability of observing by chance the same or greater correlation coefficient if the data are uncorrelated 28

2.4 PCC between NEO-FFI trait scores and the 'dark triad' scores (n=82). PP stands for primary psychopathy, SP for secondary psychopathy, M for Machiavellianism, and Nar for narcissism scores. 29

2.5 The statistics of output feature 31

2.6 The number and fraction of drug users 34

4.1 Descriptive statistics: Means, 95% CIs for means, and standard deviations for the whole sample, for non-users of illicit drugs and for users of illicit drugs. Dimensionless score z (4.1.1) for separation of users from non-users of illicit drugs is presented as well as the sensitivity and specificity P of the best separation of normal distributions with this z. Sensitivity (Sn) and Specificity (Sp) are calculated for all one-feature classifiers. Θ is the threshold for class separation: one class is given by the inequality score$\leq \Theta$ and another class by score$> \Theta$. 72

4.2 Significance of differences of means for total sample, users and nonusers of illicit drugs (p-values). 74

4.3 PCC for NEO-FFI-R for raw data 75

4.4 Polychoric correlation coefficients (PoCC) of measured psychological traits (n=1885). 75

4.5 Mean values of five factors for the three 'normal' samples and for the data. N-u, Illicit stands for non-users of illicit drugs with decade-based definition of users (they either never used illicit drugs or used them more than a decade ago), Samp stands for the total sample, U Illicit stands for users of illicit drugs for decade-based definition of users; compare with Table 4.1 80

4.6 Deviation of T-$score_{sample}$ from the sample mean for various groups of users for the decade-based user/non-user separation 83

4.7 Deviation of T-$score_{sample}$ from the sample mean for various groups of users for the year-based user/non-user separation 84

4.8 Deviation of T-$score_{sample}$ from the sample mean for various groups of users for the month-based user/non-user separation 84

4.9 Deviation of T-$score_{sample}$ from the sample mean for various groups of users for the week-based user/non-user separation 85

4.10 Significant differences of means for groups of users and non-users for the decade-based user/non-user separation. 87

4.11 Significant differences of means for groups of users and non-users for the year-based user/non-user separation. 90

4.12 Significant differences of means for groups of users and non-users for the month-based user/non-user separation. 90

4.13 Significant differences of means for groups of users and non-users for the week-based user/non-user separation. 91

4.14 The results of feature ranking. Data include country of residence and ethnicity quantified by CatPCA. FVE is the fraction of explained variance. CFVE is the cumulative FVE. The least informative features are located towards the bottom of the table. 96

4.15 The results of feature ranking. Data include dummy coded country of residence and ethnicity. FVE is the fraction of explained variance. CFVE is the cumulative FVE. The least informative features are lower located. 97

4.16 The result of sparse PCA feature ranking. Data include country of residence and ethnicity quantified by CatPCA. 98

4.17 The result of sparse PCA feature ranking. Data include dummy coded country of residence and ethnicity. 98

4.18 The best results of the drug users classifiers (decade-based definition of users). Symbol 'X' means the used input feature. Results are calculated by LOOCV. 100

4.19 The best results of the drug users classifiers in the space of the first four principal components. Symbol 'X' means used input feature. Results are calculated by LOOCV. 103

4.20 Possible class binarization 105

4.21 The best classifiers to each class binarization for ecstasy usage in the space of original input features. 106

4.22 The best class binarization in space of original attributes with the best classifiers for alcohol and cannabis. Sn is sensitivity. Sp is specificity. 108

4.23 The best class binarization in space of principal components with the best classifiers for alcohol and cannabis. Sn is sensitivity. Sp is specificity. 108

4.24 The best class binarization and the best space with the best classifier. Sn is sensitivity. Sp is specificity. Clas. means the best classifier. 109

4.25 Number of drug users for pleiades in the database 112

4.26 Statistically significant differences of means for groups of users and non-users for each pleiad for decade- year-, month-, and week-based classification problem. The symbol ' ⇓ ' corresponds to a significant difference where the mean in the users group is less than the mean in non-users group, and the symbol '⇑' corresponds to a significant difference where the mean in users group is greater than the mean in non-users group. Empty cells corresponds to insignificant differences. The difference is considered to be significant if the p-value is less than 0.01). 113

4.27 The best results of the pleiad users classifiers. Symbol 'X' means input feature used in the best classifier. Sensitivity and Specificity were calculated by LOOCV. 115

4.28 Coefficients of linear discriminant for ecstasy user/non-user separation and decade-, year-, month-, and week-based definition of users (10 attributes) 120

4.29 Performance and stability of linear discriminant for ecstasy user/non-user separation and decade-, year-, month-, and week-based definition of users (10 attributes). All indicators are in %. 121

4.30 Coefficients of linear discriminant for heroin user/non-user separation and decade-, year-, month-, and week-based definition of users (10 attributes) 121

4.31 Performance and stability of linear discriminant for heroin user/non-user separation and decade-, year-, month-, and week-based definition of users (10 attributes). All indicators are in %. 121

4.32 Coefficients of linear discriminant for ecstasy user/non-user separation and decade-, year-, month-, and week-based definition of users. (7 attributes) 123

4.33 Coefficients of linear discriminant for heroin user/non-user separation and decade-, year-, month-, and week-based definition of users. (7 attributes) 123

4.34 Means and standard deviations for users of ecstasy, for users of heroin, and for users of ecstasy OR heroin. Dimensionless z-score (4.1.1) for separation of ecstasy users from heroin users is presented as well as $P = \phi(z)$, TER, and THR. 128

4.35 Coefficients (Coeff.) of linear discriminant for separation of ecstasy users from heroin users for month-based definition of users (7 attributes). TER=70.5% THR=73.0% (the sample of ecstasy AND heroin users); TER=69.6% THR=62.2% (LOOCV). 128

4.36 Coefficients (Coeff.) of linear discriminant for separation of ecstasy users from heroin users for month-based definition of users (10 attributes). TER=75.0% THR=73.0% (the sample of ecstasy AND heroin users); TER=71.6% THR=64.9% (LOOCV). 129

4.37 The numbers of false positive (FP) and false negative (FN) errors for ecstasy user/non-user decision tree classifiers (decade-based definition of users) with linear discriminant at each node and with four different criteria of threshold selection: Accuracy, Sp+Sn, Balance (Sn=Sp), and Information Gain (IG). 132

4.38 Performance of kNN user/non-user classifiers for ecstasy (decade-based definition of users) for different k and for the standard Euclidean distance. 132

5.1 Comparison time costs of three feature selection methods 152

5.2 Protocol of FFS by LR for heroin consumption, column '#' contains number of used features, 'X' means used input feature 158

5.3 Protocol of FFS by LR for heroin consumption without features E and SS, column '#' contains number of used features, 'X' means used input feature 158

5.4 Protocol of FFS by LR for heroin consumption without features Edu, E, N, O, and SS, column '#' contains number of used features, 'X' means used input feature 159

5.5 Protocol of FFS by LR for heroin consumption with features A and Gndr, column '#' contains number of used features, 'X' means used input feature 159

5.6 Protocol of BFS by LR for heroin consumption, column '#' contains number of used features, 'X' means used input feature 160

5.7 Protocol of BFS by LR for heroin consumption without features N and SS, column '#' contains number of used features, 'X' means used input feature 161

5.8 Protocol of BFS by LR for heroin consumption without features N, SS, Edu, Imp, and O, column '#' contains number of used features, 'X' means used input feature 161

5.9 Protocol of BFS by LR for heroin consumption with features E and Gndr, column '#' contains number of used features, 'X' means used input feature 161

5.10 The results of LR models selected by ES, FFS, and BFS and AASA for heroin consumption, column '#' contains number of used features 164

5.11 The results of LR models selected by ES, FFS, and BFS and AASA for ecstasy consumption, column '#' contains number of used features 166

5.12 The results of LR models selected by ES, FFS, and BFS and AASA for cannabis consumption, column '#' contains number of used features . . . 168

5.13 First (minimal) feature set based on FFS for ecstasy usage by several classifiers, column '#' contains number of used features, 'X' means used input feature . . . 169

5.14 Second AAS of first kind based on FFS for ecstasy usage without age and O for several classifiers, column '#' contains number of used features, 'X' means used input feature . . . 169

5.15 Third AAS of first kind based on FFS for ecstasy usage without age, SS, O, and Edu for several classifiers, column '#' contains number of used features, 'X' means used input feature . . . 170

5.16 Found FFS minimal sets for features without age for each classifier for ecstasy consumption, column '#' contains number of used features, 'X' means used input feature; non-minimal set is highlighted by blue background . . . 170

5.17 Found FFS minimal sets for features without O for each classifier for ecstasy consumption; column '#' contains number of used features, 'X' means used input feature; non-minimal set is highlighted by blue background . . . 171

5.18 First (minimal) feature set based on FFS for ecstasy usage by several classifiers, column '#' contains number of used features, 'X' means used input feature . . . 171

5.19 First alternative of first kind based on BFS for ecstasy usage without age and SS by several classifiers, 'X' means used input feature . 172

5.20 Found BFS minimal sets for features without age for each classifier for ecstasy consumption, column '#' contains number of used features, 'X' means used input feature; non-minimal set is highlighted by blue background 173

5.21 Found BFS minimal sets for features without SS for each classifier for ecstasy consumption, column '#' contains number of used features, 'X' means used input feature; non-minimal set is highlighted by blue background 173

5.22 The results of FS models based on ES, FFS, and BFS and AASA results for ecstasy consumption by several classifier, column '#' contains number of used features, LW is linear regression used to find weights. 175

5.23 The results of FS models based on ES, FFS, and BFS and AASA results for cannabis consumption by several classifier, column '#' contains number of used features, LW is linear regression used to find weights. 176

5.24 The results of LR models selected by ES, FFS, and BFS and AASA results for president elections, column '#' contains number of used features, LW is linear regression used to find weights. 178

5.25 The results of FS models based on ES, FFS, and BFS and AASA results for USA president elections by several classifier, column '#' contains number of used features, LW is linear regression used to find weights. 180

5.26 Numbers of features of Breast cancer dataset 181

5.27 The results of FS models based on ES, FFS, and BFS and AASA results for Breast Cancer by several classifier, column '#' contains number of used features, LW is linear regression used to find weights. 182

5.28 The results of FS models based on ES, FFS, and BFS and AASA results for Breast Cancer by several classifier, column '#' contains number of used features, LW is linear regression used to find weights. 185

B.1 List of abbreviations 212

C.1 Mean T-$score_{sample}$ (MT) and 95% CI for it for groups of users and non-users with decade-based definition of users 230

C.2 Mean T-$score_{sample}$ (MT) and 95% CI for it for groups of users and non-users with year-based definition of users 238

C.3 Mean T-$score_{sample}$ (MT) and 95% CI for it for groups of users and non-users with month-based definition of users 246

C.4 Mean T-$score_{sample}$ (MT) and 95% CI for it for groups of users and non-users with week-based definition of users 254

C.5 PCCs between drug consumptions with decade-based user/non-user separation. Amph. stays for amphetamines and Benz. for Benzodiazepines. 263

C.6 PCCs between drug consumptions with year-based user/non-user separation. Amph. stays for amphetamines and Benz. for Benzodiazepines. 264

C.7 Coefficients of linear discriminant for user/non-user separation and decade-based definition of users (10 attributes) 267

C.8 Coefficients of linear discriminant for user/non-user separation and year-based definition of users (10 attributes) 268

C.9 Coefficients of linear discriminant for user/non-user separation and month-based definition of users (10 attributes) 269

C.10 Coefficients of linear discriminant for user/non-user separation and week-based definition of users (10 attributes) 270

C.11 Performance and stability of linear discriminant for decade-based definition of users (10 attributes). 271

C.12 Performance and stability of linear discriminant for year-based definition of users (10 attributes). 272

C.13 Performance and stability of linear discriminant for month-based definition of users (10 attributes). 273

C.14 Performance and stability of linear discriminant for week-based definition of users (10 attributes). 274

C.15 Coefficients of linear discriminant for user/non-user separation and decade-based definition of users (7 attributes) 275

C.16 Coefficients of linear discriminant for user/non-user separation and year-based definition of users (7 attributes) 276

C.17 Coefficients of linear discriminant for user/non-user separation and month-based definition of users (7 attributes) 277

C.18 Coefficients of linear discriminant for user/non-user separation and week-based definition of users (7 attributes) 278

C.19 Performance and stability of linear discriminant for decade-based definition of users (7 attributes). 279

C.20 Performance and stability of linear discriminant for year-based definition of users (7 attributes). 280

C.21 Performance and stability of linear discriminant for month-based definition of users (7 attributes). 281

C.22 Performance and stability of linear discriminant for week-based definition of users (7 attributes). 282

D.1 Minimal feature sets for heroin consumption based on ES selected by LR 284

D.2 Minimal feature sets for cannabis consumption based on ES selected by LR . . . 285

D.3 Minimal feature sets for ecstasy consumption based on ES selected by LR . . . 285

D.5 Minimal feature sets for ecstasy consumption based on ES selected by several classifiers . . . 286

D.4 Minimal sets for cannabis consumption based on ES selected by several classifiers . . . 290

D.6 P-Party Victories . . . 292

D.7 O-Party Victories . . . 293

D.8 Minimal sets for president elections of USA based on ES selected by LR . . . 293

D.9 Minimal feature sets for president elections based on ES by several classifiers . . . 294

List of Figures

2.1 Categories of drug users. Categories with green background always correspond to drug non-users. Four different definitions of drug users are presented. 32

2.2 Classes of drug users. 32

3.1 Pearson's illustration of PCA definition [92]. P_i are data points, p_i are their distances to the approximating line. The best approximation problem is UV$= \sum_i p_i^2 \to min$ 39

3.2 Nominal feature 'Country' quantification 41

3.3 Elementary branching in a decision tree 50

4.1 The distributions of SS for users and non-users of illicit drugs (normalized to 100% in each group) for the decade-based user/non-user separation. The optimal threshold is $\Theta = 4$ 73

4.2 The histograms of the number of users: A - for the decade-based user/non-user separation, B - for the month-based user/non-user separation 76

4.3 Distribution of drug usage: A: Alcohol, B: Amphetamines, C: Amyl nitrite, D: Benzodiazepines, E: Cannabis, F: Chocolate, G: Cocaine, H: Caffeine, I: Crack, and J: Ecstasy 77

4.4 Distribution of drug usage: A: Heroin, B: Ketamine, C: Legal highs, D: LSD, E: Methadone, F: Magic mushrooms, G: Nicotine, and H: VSA 78

4.5 Mean T-score NEO-FFI-R for the total sample and for non-users of illicit drugs with respect to the BLSA mean as a norm. 80

4.6 Average personality profiles for the decade-based user/non-user separation. T-scores with respect to the population norm mean (left column) and T-score$_{sample}$ with respect to the sample means (right column) for: A & B: Alcohol, C & D: LSD, E & F: Cannabis, and G & H: Heroin 88

4.7 Average personality profiles for Ketamine for the year-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means 89

4.8 Average personality profiles for Amyl nitrite for the month-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means 89

4.9 Average personality profiles for Crack for the week-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means 89

4.10 Strong drug usage correlations: A: for the decade-based classification problem and B: for the year-based classification problem 93

4.11 Pairs of decade-based drug usages with high RIG: A: approximately symmetric RIG and B: significantly asymmetric RIG. In figure B the arrow from cocaine usage to heroin usage, for example, means that knowledge of cocaine usage can decrease uncertainty in knowledge about heroin usage. 95

4.12 Conditional distribution for gender and alcohol. 100

4.13 Conditional distribution for gender and caffeine. 101

4.14 Conditional distribution for gender and chocolate. 101

4.15 Conditional distribution for gender and nicotine. 101

4.16 Decision tree for ecstasy. Input features are: Age, SS, and Gndr. Non-terminal nodes are depicted with dashed border. Values of Age, SS, and Gndr are calculated by quantification procedures described in the 'Input feature transformation' Section. The weight of each case of users class is 1.15 and of non-users class is 1. Column 'Weighted' records normalized weights: the weight of each class is divided by sum of weights. 104

4.17 The distribution of VSA users in space of the two first principal components: A) Never used, B) Used over a decade ago, C) Used in last decade, D) Used in last year, E) Used in last month, F) Used in last week, G) Used in last day, H) legend. 107

4.18 Correlation pleiades for drug use (in a circle, in a triangle and in a rectangle). Additionally, a highly correlated 'smoking couple', cannabis and nicotine, is separated by an ellipse. E stands for ecstasy, H for heroin, B for benzodiazepines, and MM for magic mushrooms. Other drugs are denoted by the first two letters of their names. Edges represent correlations. 111

4.19 Average personality profiles for HeroinPl for the decade-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means. 113

4.20 Average personality profiles for EcstasyPl for the month-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means. 114

4.21 Average personality profiles for BenzoPl for the month-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means. 114

4.22 Average personality profiles for HeroinPl for the week-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means. 114

4.23 Angles between directions of linear discriminants for user/non-user classification for ecstasy and heroin. Two types of angels are presented: between the discriminant directions for all periods and the discriminant vector for decade-based definitions of users for both drugs and angles between directions of linear discriminants for ecstasy and heroin (and the same periods). For convenience, both cosines of angles (A, B) and angles in grads (C, D) are presented. 124

4.24 Venn diagrams of relations between ecstasy and heroin use for decade-, year-, month-, and week-based definitions of users. 125

4.25 Distribution of ecstasy (NOT heroin) users, heroin (NOT ecstasy) users, and heroin AND ecstasy users in various age and education groups. 127

4.26 Mean values with their 95% confidence intervals for significantly different psychological traits of ecstasy and heroin users: A) N, B) E, C) A, D) Imp 129

4.27 A two-level classification tree for ecstasy users and non-users (decade-based definition of users) with linear discriminant classifiers at the nodes. A data cloud is visualised by projection on the plain of two first principal components for the root and the nodes of the first levels (above nodes). Users are represented by blue (light) circles, non-users by red (dark) circles. The ROC curves for the linear discriminants at each branching node are below the nodes. 131

4.28 Srceenshots of VidaExpert: a) Elastic map in the three-dimensional PCA view, b) Coloring of the map in internal coordinates. 133

4.29 Elastic maps and density visualisation for the database of drug users: a) Density of the data cloud visualisation on the elastic map presented in the internal coordinates, b)-d) Elastic map embedded in the 3-dimensional principal component space under various angles of view. Data points are in black. 135

4.30 Visualisation of various functions on the elastic map (internal coordinates): a) Age, b) Imp, c) SS, d) O (age and attributes apparently correlated with age on the maps). 136

4.31 Visualisation of various functions on the elastic map (internal coordinates): a) N, b) A, c) C, d) Edu 137

4.32 Visualisation of various functions on the elastic map (internal coordinates): a) E, b) Gndr. 138

4.33 Visualisation of linear discriminant classifiers on the non-linear PCA elastic map screen: a) Ecstasy, b) Heroin, c) Benzodiazepines d) The group of 'Illicit drugs'. White squares – users, black squares – non-users. Light (red) background – LDA predicts users, dark (blue) background – LDA predicts non-users. 139

4.34 Simplest examples of risk map of: ecstasy consumption for female (a) and male (b), heroin consumption for female (c) and male (d), and benzodiazepines consumption for female (e) and male (f). . . . 140

5.1 Construction of union model. 155

5.2 Construction of ensemble model. 155

5.3 Structure of ensemble model. 156

5.4 Comparisons of LR models selected by ES, FFS, and BFS and AAS results for heroin consumption. ES kind 1 Ensemble and ES kind 2 Ensemble model are the best models. 164

5.5 Venn diagram for features which are used for the best ES model, the first kind ensemble and union models and the second kind ensemble and union models for heroin consumption. 165

5.6 Comparisons of LR models selected by ES, FFS, and BFS and AAS results for ecstasy consumption. BFS Kind 1 Ensemble and ES Kind 2 Ensemble are the best model. 166

5.7 Venn diagram for features which are used for the best ES model, the first kind ensemble and union models and the second kind ensemble and union models for ecstasy consumption. 167

5.8 Comparisons of LR models selected by ES, FFS, and BFS and AAS results for USA president elections. 179

5.9 Venn diagram for features which are used for the best ES model, the first kind ensemble and union models and the second kind ensemble and union models for USA president elections. 179

5.10 Comparisons of LR models selected by FFS and BFS and AAS results for breast cancer diagnose. 183

5.11 Venn diagram for features which are used for the first kind ensemble and union models and the second kind ensemble and union models for breast cancer diagnosis. 184

A.1 Flowcharts blocks. 194

A.2 The best model selection. 195

A.3 Decision tree: quality (Dataset). 196

A.4 Node creation: create (SIN). 197

A.5 Node splitting: Split (Options). 198

A.6 Calculate criterion: Calculate criterion (Options, Feature, First). . . 199

A.7 Calculate criterion: Real Criterion(Values). 200

A.8 kNN (k nearest neighbours): quality(Dataset). 201

A.9 KNN Test (Options, Database, CIN). 202

A.10 Fisher's distance transformation Fisher transformed(SetNN, k). . . 203

A.11 Advanced distance transformation. 204

A.12 Calculate membership: Membership(SetNN, Vote). 205

A.13 PDFE (probability density function estimation): quality(Dataset). . 206

A.14 Fit PDFE: PDFE model(Options, Instances). 207

A.15 Form single model Single model(Options, Instances). 208

A.16 Test instance: Test(Options, Model, CIN). 209

A.17 PDFE one model test: ModelTest(Model, CIN,W). 210

B.1 **Blocks of flowcharts.** 211

B.2 General scheme for AASA and construct robust classifiers. 213

B.3 Feature subsets selection with AAS. 214

B.4 Select the optimal model. 215

B.5 Select the optimal model: FS, Cl, ES. 216

B.6 Select the Optimal model: FS, Cl, FFS. 217

B.7 Select the optimal model: FS, Cl, BFS. 218

B.8 Form list of first kind AASA models. 219

B.9 Search OMM (CFS, FFS, and Cl, RA). 220

B.10 Search OMM (CFS, BFS, Cl, RA). 221

B.11 Search OMM (CFS, ES, Cl, RA). 222

B.12 Form List of Appropriate Models (LAM). 223

B.13 Select minimal model from LAM. 224

B.14 Form list of second kind AASA models. 225

B.15 Selection of the optimal model among five candidates. 226

B.16 Form union model. 227

B.17 Form ensemble model. 228

CHAPTER 1

Introduction

1.1 Backgrounds and analysis of dataset

Data analysis is much more challenging than simply locating, identifying, understanding, and citing of data. It is the techniques of systematically implementing statistical and logical process of inspecting, cleaning, transforming, modelling and evaluating data with the aim of discovering useful information. The concept of data mining have been developed recently. In today's word data mining and machine learning have become a popular subjects. It is a family of computational methods that goal at collecting and analysing data, and it is a combination of computer science and statistics. Applications of data mining in scientific applications is widely employed in many areas such as prediction and forecasting, artificial intelligence, pattern recognition, financial data analysis and so on. The focus of data mining in this area is to analyse data to help understanding the nature of scientific datasets. The classical methods of supervised classification are widely used to meet data analysis challenges. Classification is a prevalent problem which encompasses several various applications. Classification is the task of learning a target function f which maps each input attribute $x \in X$ to output class labels $y \in \{1, 2, \ldots, C\}$, where X is the attribute space. Classification approaches are widely used for predicting or describing data sets with binary or categorical variables.

In this thesis, we analyse important practical problem dataset. This thesis is two-fold.

First, I participated in the support and maintenance of the project 'Personality traits and drug consumption' [1]. The real data from almost 2000 respondents have been analysed. In the description of the analysis of drug use I follow our book [1]. Many modern methods of data mining were employed for assessment of psychological predispositions to consumption of 18 different substances. The psychological project was designed by professional forensic psychologists and my role was in data analysis and risk assessment. The central problem is the search and validation of psychological predictors of consumption of different drugs. We employed many algorithms in order to analyse the predictability of user/non-user classification on the basis of psychological data: decision trees (DT) with various splitting criteria, random forests (RF), k-nearest neighbours (kNN) with various adaptive distances, linear discriminant analysis (LDA), Gaussian mixtures (GM), probability density function estimation (PDFE) by radial basis functions, logistic regression (LR), and naïve Bayes (NB) approach were applied to predict the risk of drug consumptions. I applied correlation analysis based on the Pearson's correlation coefficient and on relative information gain. Both models revealed the existence of groups of drugs with strongly correlated consumption. Three correlation pleiades were identified. An exhaustive search was performed to select the most effective subset of input features and data mining methods to classify users and non-users of each drug and pleiad. The quality of classification with sensitivity and specificity being greater than 70% for almost all classification tasks. The best results with sensitivity and specificity being greater than 75% were achieved for cannabis, crack, ecstasy, legal highs, LSD, and volatile substance abuse (VSA).

Second, several new methods and approaches to feature selection were proposed and tested on the drug consumption database and on several other publicly available databases. These methods include 'double Kaiser selection' for selection of

the main principal components and main attributes. Each principal component is a combination of attributes, each attribute can be presented as a distribution on factors. This symmetry allowed us to apply any heuristics invented for factors selection to feature selection as well. We developed a methodology for the selection of alternative sets of attributes of several kinds. This methodology creates and utilises controllable multicollinearity. Multicollinearity is, at the same time, a useful and an undesirable property of data. It can be useful because it allows to correct mistakes in data and to evaluate missed data. It is undesirable because many statistical tasks become ill-conditional. We propose to optimise the multicollinearity by creation of the so-called alternative attribute sets. Alternative attribute sets approach (AASA) can determine several sets of relevant attributes that can be used to solve original problems separately. AASA was built on base of minimal feature set. We tested AASA on several classification problems. We demonstrated that this methodology could be more accurate than the best traditional feature selection methods, such as exhaustive search, forward and backward feature selection. We applied AASA for three different database: 'Drug consumption' (psychology, [2,3], section 'Database' of chapter 2), 'USA president elections' (politics, [4]), and 'Breast cancer' (medicine, [5]) (see 'Database' section and 'Data sets for AASA testing' section).

1.2 Definitions of drugs and drug usage

Since Popper, it has become a commonplace opinion in the philosophy of science that the 'value' of definitions, besides in mathematics, is generally low. Nevertheless, for many more practical spheres of activity, from jurisprudence to health planning, definitions are necessary to impose theoretical boundaries on a subject, in spite of their incompleteness and their tendency to change with time. This applies strongly to definitions of drugs and drug use.

Following the standard definitions [6]:

- A *drug* is a 'chemical that influences biological function (other than by providing nutrition or hydration)'.
- A *psychoactive drug* is a 'drug whose influence is in a part on mental functions'.
- An *abusable psychoactive drug* is a 'drug whose mental effects are sufficiently pleasant or interesting or helpful that some people choose to take it for a reason other than to relieve a specific malady'.

In this study we use the term 'drug' for abusable psychoactive drug regardless of whether it is illicit or not. While legal drugs such as sugar, alcohol and tobacco are probably responsible for far more premature death than illegal recreational drugs [7], the social and personal consequences of recreational drug use can be highly problematic [8].

Use of drugs introduces risk into a life across a broad spectrum. It constitutes an important factor for increasing risk of poor health, along with earlier mortality and morbidity, and has significant consequences for society [9, 10]. Drug consumption and addiction constitutes a serious problem globally. There are numerous *risk factors* for addiction, which are defined as any attribute, characteristic, or event in the life of an individual that increase the probability of drug consumption. A number of such attributes are correlated with initial drug use, including psychological, social, individual, environmental, and economic factors [11–13]. These factors are likewise associated with a number of personality traits [14, 15]. There is a well-known problem in the analysis of the psychological deviations associated with drug use: to distinguish the result of drug use from the the cause of it [16]. To solve this problem, we have to use *relatively constant psychological traits*. Another solution is to organise large longitudinal studies which will use the traits of the patients at the different stages of drug use (such an approach seems to be more or less impossible for a number of reasons).

1.3 Personality traits

Sir Francis Galton (1884) [17] proposed to use a dictionary as a mean for constructing description of individual differences. He selected the personality-descriptive terms and stated the problem of their interrelations. In 1934, Thurstone [18] selected 60 adjectives that are in common use for describing persons and asked each of 1300 respondents to think of a person whom he knew well and to select the adjectives that can describe this person. After studying the correlation matrix he found that *five* factors are sufficient to describe this choice.

There were many versions of five factors proposed after Thurston [19], for example:

- Surgency, agreeableness, dependability, emotional stability, and culture;
- Surgency, agreeableness, conscientiousness, emotional stability, and culture;
- Assertiveness, likeability, emotionality, in-telligence, and responsibility;
- Social adaptability, conformity, will to achieve, emotional control, and inquiring intellect;
- Assertiveness, likeability, task interest, emotionality, and intelligence;
- Extraversion, friendly compliance, will to achieve, neuroticism, and intellect;
- Power, love, work, affect, and intellect;
- Interpersonal involvement, level of socialization, self-control, emotional stability, independence.

There are also systems with different numbers of factors (three, seven, etc.). The most important three-factor systems is: extraversion, psychoticism and neuroticism.

Nowadays, after many years of research and development, psychologists have largely agreed that the personality traits of the modern Five Factor Model (FFM) constitute the most comprehensive and adaptable system for understanding human individual differences [20]. The FFM comprises Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). The five traits can be summarized as:

N *Neuroticism* is a long-term tendency to experience negative emotions such as nervousness, tension, anxiety and depression (associated adjectives [21]: anxious, self-pitying, tense, touchy, unstable, and worrying);

E *Extraversion* is manifested in outgoing, warm, active, assertive, talkative, cheerful characters, often in search of stimulation (associated adjectives: active, assertive, energetic, enthusiastic, outgoing, and talkative);

O *Openness to experience* is a general appreciation for art, unusual ideas, and imaginative, creative, unconventional, and wide interests (associated adjectives: artistic, curious, imaginative, insightful, original, and wide interest);

A *Agreeableness* is a dimension of interpersonal relations, characterized by altruism, trust, modesty, kindness, compassion and cooperativeness (associated adjectives: appreciative, forgiving, generous, kind, sympathetic, and trusting);

C *Conscientiousness* is a tendency to be organized and dependable, strong-willed, persistent, reliable, and efficient (associated adjectives: efficient, organised, planful, reliable, responsible, and thorough).

1.4 The problem of relations between personality traits and drug consumption

A number of studies have illustrated that personality traits are associated with drug consumption. Roncero et al [22] highlighted the importance of the relationship between high N and the presence of psychotic symptoms following cocaine-induced drug consumption. Vollrath & Torgersen [23] observed that the personality traits of N, E, and C are highly correlated with hazardous health behaviours. A low score of C, and high score of E or high score of N correlate strongly with multiple risky health behaviours. Flory et al [24] found alcohol use to be associated with lower A and C, and higher E. They also found that lower A and C, and higher O are associated with marijuana use. Sutina et al [10] demonstrated that the relationship between low C and drug consumption is moderated by poverty; low C is a stronger risk factor for illicit drug usage among those with relatively higher socioeconomic status. They found that high N, and low A and C are associated with higher risk of drug use (including cocaine, crack, morphine, codeine, and heroin). It should be mentioned that high N is positively associated with many other addictions like internet addiction, exercise addiction, compulsive buying, and study addiction [25].

An individual's personality profile plays a role in becoming a drug user. Terracciano et al [26] demonstrated that compared to never smokers, current cigarette smokers are lower on C and higher on N. They found that profile of cocaine/ heroin users have score very high on N and very low on C and marijuana users score high on O but low on A and C. Turiano et al [27] found a positive correlation between N and O, and drug use, while, increasing scores for C and A decreases risk of drug use. Previous studies demonstrated that participants who use drugs including alcohol and nicotine have a strong positive correlation between A and C and a strong negative correlation for each of these factors with N [28,29]. Three high-order personality traits are proposed as endophenotypes for substance use

disorders: Positive Emotionality, Negative Emotionality, and Constraint [30].

Sensation seeking is also higher for users of recreational drugs [31]. The problem of risk evaluation for individuals is much more complex. This was explored very recently by Yasnitskiy et al [32], Valeroa et al [33] and Bulut & Bucak [34]. Both individual and environmental factors predict substance use, and different patterns of interaction among these factors may have different implications [35]. Age is a very important attribute for diagnosis and prognosis of substance use disorders. In particular, early adolescent onset of substance use is a robust predictor of future substance use disorders [36].

Valeroa et al [33] evaluated the individual risk of drug consumption for alcohol, cocaine, opiates, cannabis, ecstasy and amphetamines. Input data were collected using the Spanish version of the Zuckerman-Kuhlman Personality Questionnaire (ZKPQ). Two samples were used in this study. The first one consisted of 336 drug dependent psychiatric patients of one hospital. The second sample included 486 control individuals. The authors used a decision tree as a tool to identify the most informative attributes. The sensitivity (proportion of correctly identified positives) of 40% and specificity (proportion of correctly identified negatives) of 94% were achieved for the training set. The main purpose of this research was to test if predicting drug consumption was possible and to identify the most informative attributes using data mining methods. Decision tree methods were applied to explore the differential role of personality profiles in drug consumer and control individuals. The two personality factors, Neuroticism and anxiety and the ZKPQ's Impulsivity, were found to be most relevant for drug consumption prediction. The low sensitivity (40%) score means that such a decision tree cannot be applied to real life situations.

Without focussing on specific addictions, Bulut & Bucak [34] estimated the proportion of teenagers who exhibit a high risk of addiction. The attributes were collected by an original questionnaire, which included 25 questions. The form was filled in by 671 students. The first 20 questions asked about the teenagers'

financial situation, temperament type, family and social relations, and cultural preferences. The last five questions were completed by their teachers and concerned the grade point average of the student for the previous semester according to a 5-point grading system, whether the student had been given any disciplinary punishment so far, if the student had alcohol problems, if the student smoked cigarettes or used tobacco products, and whether the student misused substances. In Bulut et al's study there are five risk classes as outputs. The authors diagnosed teenagers risk of being a drug abuser using seven types of classification algorithms: *k*-nearest neighbor, ID3 and C4.5 decision tree based algorithms, naïve Bayes classifier, naïve Bayes/decision trees hybrid approach, one-attribute-rule, and projective adaptive resonance theory. The classification accuracy of the best classifier was reported as 98%.

Yasnitskiy et al [32], attempted to evaluate the individual's risk of illicit drug consumption and to recommend the most efficient changes in the individual's social environment to reduce this risk. The input and output features were collected by an original questionnaire. The attributes consisted of: level of education, having friends who use drugs, temperament type, number of children in the family, financial situation, alcohol drinking and smoking, family relations (cases of physical, emotional and psychological abuse, level of trust and happiness in the family). There were 72 participants. A neural network model was used to evaluate the importance of attributes for diagnosis of the tendency to drug addiction. A series of virtual experiments was performed for several test patients (drug users) to evaluate how possible it is to control the propensity for drug addiction. The most effective change of social environment features was predicted for each patient. The recommended changes depended on the personal profile, and significantly varied for different patients. This approach produced individual bespoke advice to effect decreasing drug dependence.

In this study we tested associations with personality traits for different types of drugs separately, using the Revised NEO Five-Factor Inventory (NEO-FFI-

R) [37], the Barratt Impulsiveness Scale Version 11 (BIS-11) [38], and the Impulsivity Sensation-Seeking Scale (ImpSS) [39] to assess impulsivity and sensation-seeking respectively.

1.5 New dataset 'drug consumption'

In this study [2], the database was collected by an anonymous online survey methodology by Elaine Fehrman, yielding 2051 respondents. The database is available online [3]. Twelve attributes are known for each respondent: personality measurements which include N, E, O, A, and C scores from NEO-FFI-R, impulsivity (Imp) from (BIS-11), sensation seeking (SS) from (ImpSS), level of education (Edu), age, gender (Gndr), country of residence, and ethnicity. The data set contains information on the consumption of 18 central nervous system psychoactive drugs including alcohol, amphetamines, amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, magic mushrooms (MMushrooms), nicotine, and Volatile Substance Abuse (VSA). For uniformity we consider VSA as drug. One fictitious drug (Semeron) was introduced to identify over-claimers. For each drug, participants selected either: they never used this drug, used it over a decade ago, or in the last decade, year, month, week, or day.

Participants were asked about various substances, which were classified as either central nervous system depressants, stimulants, or hallucinogens. The depressant drugs comprised alcohol, amyl nitrite, benzodiazepines, tranquilizers, gamma-hydroxybutyrate solvents and inhalants, and opiates such as heroin and methadone/prescribed opiates. The stimulants consisted of amphetamines, nicotine, cocaine powder, crack cocaine, caffeine, and chocolate. Although chocolate contains caffeine, data for chocolate was measured separately, given that it may induce parallel psychopharmacological and behavioural effects in individuals congruent to other addictive substances [40]. The hallucinogens included

cannabis, ecstasy, ketamine, LSD, and magic mushrooms. Legal highs such as mephedrone, salvia, and various legal smoking mixtures were also measured.

We use four different definitions of 'drug users' based on how recent the use was. Firstly, for decade-based separation we merge two isolated categories ('Never used' and 'Used over a decade ago') into the class of non-users, and all other categories are merged to form the class of users. For year-based classification we now additionally merge the category 'Used in last decade' into the group of non-users and place four other categories ('Used in last year-month-week-day') into group of users. We continue separating into users and non-users depending on the time scale we are looking at in this nested "Russian doll" style. We also consider 'month-based' and 'week-based' user/non-user separations.

The objective of the study was to assess the potential effect of Big Five personality traits, impulsivity, sensation-seeking, and demographic data on drug consumption for different drugs, groups of drugs and for different definitions of drug users. The study had two purposes: (i) to identify the association of personality profiles (i.e. NEO-FFI-R) with drug consumption and (ii) to predict the risk of drug consumption for each individual according to their personality profile.

The sample was created by an anonymous online survey. It was found to be biased when compared with the general population, the comparison being based on the data published by Egan, et al [41] and Costa Jr & McCrae [37]. Such a bias is usual for clinical cohorts [26, 42].

1.6 First results in brief

This study reveals that the personality profiles are strongly associated with membership of groups of the users and non-users of the 18 drugs. For analysis, we use the following subdivision of the sample T-*score*: the interval 44-49 indicates a moderately low score, $(-)$, the interval 49-51 indicates a neutral score (0), and the interval 51-56 indicates a moderately high $(+)$ score. We found that the N

and O scores of drug users of all 18 drugs are moderately high $(+)$ or neutral (0), except for crack usage for the week-based classification, for which the O score is moderately low $(-)$. The A and C scores are moderately low $(-)$ or neutral (0) for all groups of drug users and all user/non-user separations. For most groups of illicit drug users the A and C scores are moderately low $(-)$ with the exception of two groups: the A score is neutral (0) in the year-based classification for LSD users and in the week-based classification for LSD and magic mushrooms users. The A and C scores for groups of legal drugs users (i.e. alcohol, chocolate, caffeine, and nicotine) are neutral (0), apart from nicotine users, whose C score is moderately low $(-)$ for all bases of user/non-user separation.

The impact of the E score is drug specific. For example, for the week-based user/non-user separation we observe:

- The E score of users is moderately low $(-)$ for amphetamines, amyl nitrite, benzodiazepines, heroin, ketamine, legal highs, methadone, and crack;
- The E score of users is moderately high $(+)$ for cocaine, ecstasy, LSD, magic mushrooms, and VSA;
- The E score of users is neutral (0) for alcohol, caffeine, chocolate, cannabis, and nicotine.

For more details see Section 4.5.

Usage of some drugs are correlated significantly. The structure of these correlations is analysed in Section 'Correlation between usage of different drugs'. Two correlation measures are utilised: the Pearson's Correlation Coefficient (PCC) and the Relative Information Gain (RIG). We found three groups of drugs with highly correlated use. The central element is clearly identified for each group. These centres are: *heroin, ecstasy, and benzodiazepines*. This means that drug consumption has a 'modular structure', which is made clear in the correlation graph. The idea of merging correlated attributes into 'modules' referred to as *correlation pleiades* is popular in biology [43–45].

The concept of correlation pleiades was introduced in biostatistics in 1931 [43]. They were used for identification of the modular structure in evolutionary physiology [43–46]. According to Berg [45], correlation pleiades are clusters of correlated traits. In our approach, we distinguish the core and the peripheral elements of correlation pleiades and allow different pleiades to have small intersections in their periphery. 'Soft' clustering algorithms relax the restriction that each data object is assigned to only one cluster (like probabilistic [47] or fuzzy [48] clustering). See the book of R. Xu and D. Wunsch [49] for a modern review of hard and soft clustering. We refer to [50] for a discussion of clustering in graphs with intersections .

The three groups of correlated drugs centered around heroin, ecstasy, and benzodiazepines are defined for the decade-, year-, month-, and week-based classifications:

- The heroin pleiad includes crack, cocaine, methadone, and heroin;
- The ecstasy pleiad consists of amphetamines, cannabis, cocaine, ketamine, LSD, magic mushrooms, legal highs, and ecstasy;
- The benzodiazepines pleiad contains methadone, amphetamines, cocaine, and benzodiazepines.

Analysis of the intersections between correlation pleiades of drugs leads to important questions and hypotheses:

- Why is cocaine a peripheral member of all pleiades?
- Why does methadone belong to the periphery of both the heroin and benzodiazepines pleiades?
- Do these intersections reflect the structure of individual drug consumption or the structure of the groups of drug consumers?

Correlation analysis of the decade-based classification demonstrates that the consumption of legal drugs (i.e. alcohol, chocolate and caffeine) is not correlated with consumption of other drugs. The consumptions of seven illicit drugs (i.e. amphetamines, cannabis, cocaine, ecstasy, legal highs, LSD, and mushrooms) are symmetrically correlated (when the correlations are measured by relative information gain, which is not symmetric a priori). There are also many strongly asymmetric correlations. For example, knowledge of amphetamines, cocaine, ecstasy, legal highs, LSD, and magic mushroom consumption is useful for the evaluation of ketamine consumption. On the other hand, knowledge of ketamine consumption is significantly less useful for the evaluation of usage of the drugs listed above.

In this study, we evaluated the individual drug consumption risk separately, for each drug and pleiad of drugs. We also analysed interrelations between the individual drug consumption risks for different drugs. We applied several data mining approaches: decision tree, random forest, k-nearest neighbours, linear discriminant analysis, Gaussian mixture, probability density function estimation, logistic regression, and naïve Bayes. The quality of classification was surprisingly high. We tested all the classifiers by *Leave-One-Out Cross Validation*. The best results, with sensitivity and specificity greater than 75%, were achieved for cannabis, crack, ecstasy, legal highs, LSD, and VSA. Sensitivity and specificity greater than 70% were achieved for the following drugs: amphetamines, amyl nitrite, benzodiazepines, chocolate, caffeine, heroin, ketamine, methadone and nicotine. The poorest result was obtained for prediction of alcohol consumption. An exhaustive search was performed to select the most effective subset of input features, and data mining methods to classify users and non-users for each drug.

Users for each correlation pleiad of drugs are defined as users of any of the drugs from the pleiad. We consider the classification problem for drug pleiades for the decade-, year-, month-, and week-based user/non-user separations. For good statistical prediction of a binary outcome it is helpful to have more or less equal

numbers of cases of each of the two classes. For pleiades sample is better balanced. For each separate drug there are too few users so we get more robust prediction at the pleiad level. For example, in the database for the week-based definition of users there are 184 users for the heroin pleiad but only 29 heroin users.

The quality of classification is high. Consider the month-based user/non-user separation of the heroin pleiad consumption. The best classifier is a decision tree with five features and sensitivity 74.18% and specificity 74.11%. A decision tree with seven attributes is the best classifier for the year-based classification problem of the ecstasy pleiad users/non-users and has sensitivity 80.65% and specificity 80.72%. In the week-based separation of the benzodiazepines pleiad users/non-users, the best classifier is a decision tree with five features and sensitivity 75.10%, and specificity 75.76%.

The creation of classifiers provided the capability to evaluate the risk of drug consumption in relation to individuals. The risk map is a useful tool for data visualisation and for the generation of hypotheses for further study (see Section 'Risk evaluation for the decade-based user/non-user separation').

1.7 The reviews of feature selection problem

In many applied data mining problems (classification, regression, etc.) the set of input attributes can be reduced. The minimal set of attributes sufficient for solution of the problem may be much smaller than initial set. Nevertheless, it may be useful to use more attributes to solve the problem when some attributes from the minimal set are unavailable or have erroneous or noise contaminated values. Therefore, some alternative sets of attributes may be useful to build a good predictor. Searching of minimal set of attributes is a Feature Selection (FS) problem. FS is a widely employed procedure for dimensionality reduction among practitioners [51]. A goal of FS is to remove irrelevant and redundant attributes

and choose an optimal subset of the relevant attributes from the original set of features [52]. The optimality feature set is defined by certain evaluation criterion, which typically leads to higher accuracy and better model interpretability. FS can improve the computational efficiency of the classification algorithm as well [51, 53–56].

In this work we consider the wrapper FS [51, 55, 57] approaches only because it is developed to select the features in accordance with accuracy of problem solution. Three widely used FS techniques are: Exhaustive Search (ES), Forward Feature Selection (FFS), and Backward Feature Selection (BFS). ES checks all possible feature subsets and it is the only way to find optimal solution. Unfortunately this approach required too much computational resources and cannot be used for more than 20 features because for set of m input features the number of possible subsets is 2^m [56]. Surrogates of ES are greedy algorithms (FFS and BFS) to find quasi optimal set of input features. Compared to BFS method, FFS method can take a quasi-optimal input feature subset with less time (See section 5.3) Recall that a goal of FS is to remove irrelevant and redundant features [54, 55]. John, Kohavi, and Pfleger [57] suggested two degrees of relevance are required:

- Strong relevance: A feature X is called strong relevant if X absolutely essential in the sense and it cannot be removed without decrease of performance.
- Weak relevant: A feature X is called weak relevant if X can sometimes provide performance.

A feature X is relevant if it is strong or weak relevant; otherwise X is irrelevant [55, 57]. Relevance of a feature does not imply optimality, while optimality does not imply relevance (see [55]). Irrelevant features provide no useful information and it should be removed because these are the sources of noise only and does not influence the aim concept in anyway. Notion of irrelevant feature is absolute for problem under consideration while the notion of redundancy is relative: since one relevant feature may be redundant in the presence of another

relevant features with which it is strongly correlated [54]. The reason of removing of redundant features is the multicollinearity problem which is a common problem for most of statistic based methods (for example, linear regression, logistic regression, k nearest neighbour, etc.).

To illustrate differences between irrelevant and redundant attributes let us consider problems of prediction of ecstasy consumption which is described in chapter 4. There are ten input attributes: age, education, N, E, O, A, C, Imp, SS, and Gndr. To evaluate risk to be ecstasy consumer with sensitivity and specificity at least 65% we can use one of 23 feature sets union of which includes all ten attributes. These sets are defined by ES for logistic regression (see Table D.3). If we apply FFS then minimal set with required accuracy includes age and O only, if we apply BFS then resulting set includes age and SS only but we can solve problems on base of any other 21 sets. It means that all attributes are relevant but each attribute is redundant at least in one feature set.

Introduction of redundancy is widely used in data transfer. For example, error-correcting codes add redundant information into message to detect and correct errors of transfer [58]. Ensemble models in machine learning also use redundancy to create model which usually over perform single classifier [59]. Recall, that we cannot use redundant attributes in one classifier but we can create several classifiers and then form ensemble model. We use this approach to improve classification accuracy.

1.8 A methodology for selection of alternative sets of attributes

In this thesis, we developed a methodology for selection of several kinds of alternative sets of attributes and usage all these sets together. We called this approach '*Alternative Attributes Sets Approach*' (AASA). AASA can find several different sets of relevant attributes such that each set can be used to solve original

problems separately. Such set with the fitted model is called Alternative Attribute Set (AAS). AAS notion is based on the notion of minimal attribute set. Minimal attribute set $M(S)$ is a subset of attributes set S which provides required accuracy and does not contain any other minimal subsets. It is necessary to stress that minimal feature set vary rare is optimal feature set. Minimal set can be found by ES, FFS, or BFS. For the minimal set selection either one or several classifier can be used. In this work for selection AAS we use six classification methods: decision trees, k-nearest neighbours, linear discriminant analysis, Gaussian mixtures, probability density function estimation, logistic regression and naïve Bayes. We defined two kinds of AAS. The first kind AAS for feature set S is the minimal set which does not contain any elements of S. The second kind AAS is the list of sets with two properties: (i) each set does not contain at least one element of set S and (ii) for each element of S there is at least one set which does not contain this element. Each first kind AAS is always the second kind AAS.

1.9 Model construction for Alternative Attributes Sets (AAS)

AASA is developed to create more robust and/or more accurate model. There are two ways to use AAS for model construction. If there are several sets which are alternative to each other (several AAS), then we can unite feature sets of all alternatives in one set and construct a model with this set of features. Let us call such model 'union model'. Another way is a creation of two level model: classifiers fitted for each AAS in the first layer and second layer contains simple classifier which used outcomes of the first level classifiers as input. We call such model 'ensemble model' [59, 60]. We consider linear ensembles only. Linear ensemble outcome is defined as $\sum w_i e_i / \sum w_i$, where e_i is outcome of ith classifier of the first layer and w_i is its weight. We use several approaches to define weights of ensemble model: simple voting, Nelder-Mead, Luus-Jaakola, pattern search, and linear

regression.

1.10 AASA applications to the datasets

We applied AASA for three real life datasets which have a different number of records and different dimensions and are taken from different application areas. The first database is 'Drug consumption' (psychology, [2], section 'Database' of chapter 2) which described more details in 'Database' section 2.1. The second database is 'USA president elections' (politics, [4]), and the third dataset is 'Breast cancer' (medicine, [5]) (see section Data sets for AASA testing).

For each classification problem and each FS method we define five candidates to the best model. The first candidate is the best model among models which are tested by basic FS method (ES, FFS, or BFS). It is not AASA model and is used as a reference point for comparison with AASA models. We call it the best FS model. The other four models are AASA models: union model for the first kind AAS, ensemble model for the first kind AAS, union model for the second kind AAS, and ensemble model for the second kind AAS (description of union model and ensemble models are presented in section 'Model construction for Alternative Attributes Sets (AAS)'). Then we select the best model among FS models and the best model among AASA models.

We found that the best AASA model is usually much better than the best FS model. In most cases the best AASA model is ensemble model for first or second kind AAS. It means that AASA can play a great role to create most accurate model because they can significantly improve accuracy. For example, for drug consumption database, ensemble model for heroin consumption for the first kind AAS for ES is much better than the best FS model. This ensemble model uses eight features and has sensitivity 71.23% and specificity 72.15% while the best FS model has sensitivity 73.11% and specificity 69.99% (see Table 5.10). The best FS model for ecstasy consumption contains seven features. This model has sensitiv-

ity 74.83% and specificity 74.52%. The best AASA model for ecstasy consumption is ensemble model for the first kind AAS for BFS. This ensemble model contains four features instead of seven and has sensitivity 74.97% and specificity 74.78% (see Table 5.11).

For USA president election database we found that four features of the 12 are enough to achieve high accuracy in prediction of the USA president elections. The best minimal model for ES and LR contains four features. This result was reproduced by several classifiers. The best ES model contains five features and has 100% sensitivity and 100% specificity. This FS model is completely included by the second kind AAS ensemble model for FFS which contains seven features and has the same accuracy.

On the other hand, for breast cancer problems ensemble model also improved accuracy. Ensemble model for the first kind AAS for FFS is better than the best FFS model, but contains 19 features instead of 9. LR ensemble model for the first kind AAS for FFS has sensitivity 98.04% and specificity 98.11%. This accuracy is much better than the 97.3% accuracy based on the best single-plane diagnostic classifier based on features mean texture, the worst area, and the worst smoothness which is obtained by [61]. Our study shows that the best LR minimal model for FFS contains 3 of the 30 features: mean concave points, worst area, and worst texture. This model has sensitivity 95.80% and specificity 96.23%. This accuracy is considerably better than 89% accuracy of model based on individual cell analysis which is obtained by [62] and than 96.2% of the LR cross validated classification accuracy which is obtained by [61] for three other features: worse radius, worse texture, and worse concave points.

1.11 Thesis Outline

We organize the thesis as follows:

- **Chapter 1**

In this chapter, we explain the notion of drug use and which personality traits we use to analyse the predisposition to use of drugs. We also review some previous pertinent results, describe the problems which we aim to solve, and briefly outline the answer. We review of feature selection problem and we study how to use redundant and alternative input features for the apply data set and analysis the classification methods of several classical benchmarks and briefly outline the answer.

- **Chapter 2**

  In this chapter, we present data, describing the attributes we measure, and the method of data collection. The Five factors model of personality is introduced: Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. The four different definitions of drug users we use are based on recency of use: decade-based, year-based, month-based, or week-based definitions. The numbers of users of 18 different psychoactive substances for the four definitions of users in the database are presented. We present two other available database used for AASA testing.

- **Chapter 3**

  In this chapter, we review the used methods of data analysis, from elementary T-scores to non-linear Principal Component analysis (PCA), including polychoric correlation, nonlinear CatPCA (Categorical Principal Component Analysis), sparse PCA, method of principal variables, original 'double' Kaiser selection rule, k Nearest Neighbours for various distances, decision tree with three split criteria (information gain, Gini gain or DKM gain), linear discriminant analysis, Gaussian mixture, probability density function estimation by radial-basis functions, logistic regression, naïve Bayes approach, random forest, and criteria for selecting the best method. We briefly identify how to use alternative attributes sets which can be useful to build a good predictor. We review the used techniques to define weights of ensemble models such as Nelder-Mead, Luus-Jaakola, Pattern search, and Linear

Regression.

- **Chapter 4**

  In this chapter, we evaluate the individual drug consumption risk separately, for each drug and pleiad of drugs. We also analyse interrelations between the individual drug consumption risks for different drugs. We use several classification methods to predict the risk of drug consumption for individuals. The objective of this chapter is to assess the potential effect of personality traits, impulsivity, sensation-seeking, and demographic data on drug consumption. This chapter has two purposes. Firstly, to identify the association of personality profiles (i.e. NEO-FFI-R) with drug consumption. Secondly, to predict the risk of drug consumption for each individual according to their personality profiles. We found the best classifiers and most significant predictors for each individual drug. We found three groups of drugs with highly correlated use and introduce three correlation pleiades. For each pleiad, the central element is clearly identified. We evaluate the individual risk of drug consumption separately, for pleiad of drugs. The best robust classifiers and most significant predictors are found for use of pleiades of drugs. We illustrate all the analysis and results by many possible ways to solve drug consumption problems. The risk map technology is developed for the visualization of the probability of drug consumption.

- **Chapter 5**

  In this chapter, we introduce the concept of creation of more robust classifier on base of AASA. We study how to use redundant and alternative input features for applied data set and analyse many classification methods of several classical benchmarks. We develop a methodology for selection of several kinds of alternative sets of attributes and usage of all these sets together. AASA can find several different sets of relevant attributes such that each set can be used to solve original problem separately. Such set with fitted model is called AAS. AAS notion is based on the notion of minimal

attribute set. In this study for selection AAS we use the six classifiers.

- **Chapter 6**

  This is the final chapter and it discusses all the results presented in this thesis.

  The results have been presented at the following conferences:

  1. Conference of the International Federation of Classification Societies, University of Bologna, 8th July 2015. The Five Factor Model of personality and evaluation of drug consumption risk.
  2. European Conference on Data Analysis, University of Essex, 2nd September 2015. Evaluation of Risk of Drug Consumption.

  and partially published in

  1. Fehrman E, Muhammad AK, Mirkes EM, Egan V, Gorban AN. The Five Factor Model of personality and evaluation of drug consumption risk. ArXiv preprint arXiv, 2015. https://arxiv.org/abs/1506.06297.
  2. Fehrman E, Muhammad AK, Mirkes EM, Egan V, Gorban AN. The Five Factor Model of personality and evaluation of drug consumption risk. In Francesco Palumbo et al. (eds.), Data Science, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 2017, pp. 215–226. http://doi.org/10.1007/978-3-319-55723-6_18.
  3. Fehrman E, ..., Muhammad AK. Personality Traits and Drug Consumption: A Story Told by Data. Springer, 2018, to be published (accepted).

CHAPTER 2

Data

2.1 Database

In this study the main database used was collected by Elaine Fehrman between March 2011 and March 2012. In January 2011, the research proposal was approved by the University of Leicester's Forensic Psychology Ethical Advisory Group, and subsequently received strong support from the University of Leicester School of Psychology's Research Ethics Committee (PREC).

The data are available online [3]. An online survey tool from Survey Gizmo [63, 64] was employed to gather data which maximised 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 a 12-month recruitment period. Of these people, 166 did not respond correctly to a validity check built into the middle of the scale, so were presumed to be inattentive to the questions being asked. Nine of these were also found to have endorsed use of a fictitious recreational drug, which was included precisely to identify respondents who over-claim, as have other studies of this kind [65]. This led a useable sample of 1885 participants (male/female = 943/942).

The snowball sampling methodology was implemented. Snowball sampling may be described as a special technique for finding research respondents [66]. One subject links the researcher to another subject, who in turn provides the link to a third, and so on.

This strategy is a tool to overcoming the problems associated with sampling concealed populations such as the criminal and the isolated. Snowball sampling belongs to a wider set of link-tracing methodologies which aims to utilize the social networks of identified respondents to provide a researcher with an ever-expanding set of potential contacts. This approach is based on the assumption that a 'path' of links exists between the initial sample and others in the same target group.

Snowball sampling is used for two primary purposes. Firstly, as an 'informal' method to reach a target population. Secondly, it is applied as a more formal methodology for making inferences about a population of individuals who have been difficult to enumerate through by the classical survey methods. This approach has enabled access to previously hidden populations. Members of such populations may be involved in activities that are considered deviant, such as drug taking, making them reluctant to take part in more formalised studies.

Snowball samples have well-known deficiencies, the principle one being that of representativeness. Most snowball samples are biased because the participants are dependent on the subjective choices of the respondents first accessed and on the structure of their links. This problem may be partially resolved, through the generation of large samples and by the replication of results ('restarting' of snowballs). Statistical analysis of snowball samples is widely studied [67].

Snowball sampling is now a standard tool for analysis of drug-using populations [68–70]. In this study, this methodology recruited a primarily (93.7%) 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 (20; 1.1%). A total of 118 (6.3%) came from a diversity of other countries,

Table 2.1. Country of residence and ethnicity composition

Country of residence			Ethnicity		
Country	#	Fraction	Cultural background	#	Fraction
Australia	54	2.9%	Asian	26	1.4%
Canada	87	4.6%	Black	33	1.8%
New Zealand	5	0.3%	Mixed- Black Asian	3	0.2%
Other (please state)	118	6.3%	Mixed- White Asian	20	1.1%
Republic of Ireland	20	1.1%	Mixed- White Black	20	1.1%
UK	1044	55.4%	Other (please state)	63	3.3%
USA	557	29.5%	White	1720	91.2%

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 was the 18 to 24 range, some 40% of the cohort was 35 or above, which is an age range often missed in studies of this kind. Table 2.1 shows country of residence and ethnicity composition.

The sample recruited was highly educated, with just under two thirds (59.5%) educated to, at least, degree or professional certificate level: 14.4% (271) reported holding a professional certificate or diploma, 25.5% (481) an undergraduate degree, 15% (284) a master's degree, and 4.7% (89) a doctorate. Approximately 26.8% (506) of the sample had received some college or university tuition although they did not hold any certificates; lastly, 13.6% (257) 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 (See Table 2.2).

Table 2.2. The age bands and education level

Age			Education		
Age band	Cases	Fraction	Education level	Cases	Fraction
18-24	643	34.1%	Left school before 16 years	28	1.5%
25-34	481	25.5%	Left school at 16 years	99	5.3%
35-44	356	18.9%	Left school at 17 years	30	1.6%
45-54	294	15.6%	Left school at 18 years	100	5.3%
55-64	93	4.9%	Some college or university, no certificate or degree	506	26.8%
65+	18	1.0%	Professional certificate/ diploma	271	14.3%
			University degree	480	25.5%
			Master's degree	284	15.0%
			Doctorate degree	89	4.7%

2.2 Personality measurements

In order to assess the personality traits of the sample, the Revised NEO Five-Factor Inventory (NEO-FFI-R) questionnaire was employed [20]. 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) [71]. 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) [20]. The five factors are: N, E, O, A, and C with 12 items per domain.

All of these domains are hierarchically defined by specific facets [72]. Egan et al [41] observe that the score for the O and E domains of the NEO-FFI instrument are less reliable than for N, A, and C. The personality traits are far from being independent. They are correlated, with higher N being associated with lower E, lower A and lower C, and higher E being associated with higher C (see Table 2.3 for more details).

In this study, participants were asked to read the 60 NEO-FFI-R statements and indicate on a five-point Likert–type 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 would be associated with high N, and low A and C. Symbolically, we depict this profile as N$\Uparrow$, A$\Downarrow$ and C$\Downarrow$. This combination was

Table 2.3. Pearson's correlation coefficients (PCC) between NEO-FFI trait scores for a large British sample, $n = 1025$ [41]; the p-value is the probability of observing by chance the same or greater correlation coefficient if the data are uncorrelated

	N	E	O	A	C
N		−0.40**	0.07*	−0.22**	−0.36**
E	−0.40**		0.16**	0.22**	0.30**
O	0.07*	0.16**		0.08*	−0.15**
A	−0.22**	0.22**	0.08*		0.13**
C	−0.36**	0.30*	−0.15**	0.13**	

$^{*}p < 0.02$, $^{**}p < 0.001$

observed for various types of psychopathy and deviant behavior, for example, in analysis of the *'dark triad' of personality*: Machiavellianism, Narcissism and Psychopathy [73].

- *Machiavellianism* refers to interpersonal strategies that advocate self-interest, deception and manipulation. A questionnaire MACH-IV is now the most widely used tool to measure MACH, the score of Machiavellianism [74]. Persons high in MACH are likely to exploit others and less likely to be concerned about other people beyond their own self-interest.

- The concept of *narcissism* comes from the Greek myth of Narcissus who falls in love with this own reflection. Formalised in psychodynamic theory, it describes a pathological form of self-love. One commonly used operational definition of *Narcissism* is based on the Narcissistic Personality Inventory, that measures persistent attention seeking, extreme vanity, excessive self-focus, and exploitativeness in interpersonal relationships. It comprises four factors: Exploitativeness/Entitlement, Leadership/Authority, Superiority/Arrogance and Self-Absorption/Self-Admiration [75].

- Levenson's self-report measure of psychopathy measure two facets of psychopathy. Factor 1 reflects *primary psychopathy* (e.g., selfishness, callousness, lack of interpersonal affect, superficial charm and remorselessness), and Factor 2 measures anti-social lifestyle and behaviours, and is akin to *secondary psychopathy* (excessive risk-takers who exhibit usual amounts of

Table 2.4. PCC between NEO-FFI trait scores and the 'dark triad' scores (n=82). PP stands for primary psychopathy, SP for secondary psychopathy, M for Machiavellianism, and Nar for narcissism scores.

	N	E	O	A	C
PP	0.30***	0.08	−0.21*	−0.43***	−0.21*
SP	0.47***	0.04	−0.21*	−0.23**	−0.19*
M	0.38***	−0.13	−0.17	−0.41***	−0.27**
Nar	−0.10	0.10	0.10	−0.43***	−0.24**

$^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$

stress and guilt).

Analysis of the 'dark triad' of personality exclaim showed [73] that N was positively associated with psychopathy and Machiavellianism (Table 2.4), the dark dimension of personality can be described in terms of low A, whereas much of the anti-social behaviour in normal people appears underpinned by high N and low C.

The so-called '*negative urgency*' is the tendency to act rashly when distressed, and is also characterized by high N, low C, and low A [76]. Negative urgency predicted alcohol dependence symptoms in disordered women, drinking problems and smoker status in pre-adolescents, and aggression, risky sex, illegal drug use, drinking problems, and disordered behavior in college students.

Thus, the hypothesis about the personality profile N⇑, A⇓ and C⇓ for drug users has solid background. We validated this hypothesis on the data and found, indeed, that for some groups of drug users it holds true. For example, for heroin and methadone users we found this classical combination. At the same time, we found various deviations from this profile for other drugs. For example, for groups of recent LSD users (used less than a year ago, or used less than a month ago, or used less than a week ago) N does not deviate significantly from the mean but O and C do: O⇑, C⇓. In this work we suggest also that O is higher for many groups of drug users. Detailed profiles of all groups of users are presented in Appendix C.

The second measure used was the Barratt Impulsiveness Scale (BIS-11) [38]. 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 [77]. The scale's items are scored on a four-point Likert-type scale. This study modified the response range to make it compatible with previous related studies [78]. 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 the BIS-11 scores are, the higher the impulsivity level [79] is. 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 [38,77]).

Impulsivity has been shown to predict aggression and heavy drinking [80]. Poor social problem solving has been identified as a potential mediating variable between impulsivity and aggression. It is likely that the cognitive and behavioural features of impulsivity militate against the acquisition of good social problem-solving skills early in life and that these deficits persist into adulthood, increasing the likelihood of interpersonal problems.

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 [39]. 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 [81].

We call this dataset 'Drug consumption'(psychology, which is presented in [2,3]). 'Drug consumption' dataset is used as one of the dataset for AASA applications.

Table 2.5. The statistics of output feature

Targets	Original categories						
	Never used	Used over a decade ago	Used in last decade	Used in last year	Used in last month	Used in last week	Used in last day
Alcohol	34	34	68	198	287	759	505
Amphetamines	976	230	243	198	75	61	102
Amyl nitrite	1305	210	237	92	24	14	3
Benzodiazepines	1000	116	234	236	120	84	95
Cannabis	413	207	266	211	140	185	463
Chocolate	32	3	10	54	296	683	807
Cocaine	1038	160	270	258	99	41	19
Caffeine	27	10	24	60	106	273	1385
Crack	1627	67	112	59	9	9	2
Ecstasy	1021	113	234	277	156	63	21
Heroin	1605	68	94	65	24	16	13
Ketamine	1490	45	142	129	42	33	4
Legal highs	1094	29	198	323	110	64	67
LSD	1069	259	177	214	97	56	13
Methadone	1429	39	97	149	50	48	73
MMushrooms	982	209	260	275	115	40	4
Nicotine	428	193	204	185	108	157	610
VSA	1455	200	135	61	13	14	7
Semeron	1877	2	3	2	1	0	0

2.3 Drugs and categories of users

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, magic mushrooms, nicotine and VSA) and one fictitious drug (semeron) which was introduced to identify over-claimers. There are 19 target features (output feature). Four target features correspond to legal drugs: alcohol, caffeine, nicotine, and chocolate. Semeron is non-existent drug and all results for this drug can be interpreted as prediction the tendency to over-claim. Statistic of all 19 target features is presented in Table 2.5. The non-existent drug semeron has eight users. This value is less than 0.5% of total number of records. Since drug semeron is not existing, availability of users provide us to estimate level of lie as 0.5%. We exclude semeron for further study by this reason. It means, for further study we have to use 18 drugs.

It was recognised at the outset of this study that drug use research regularly (and

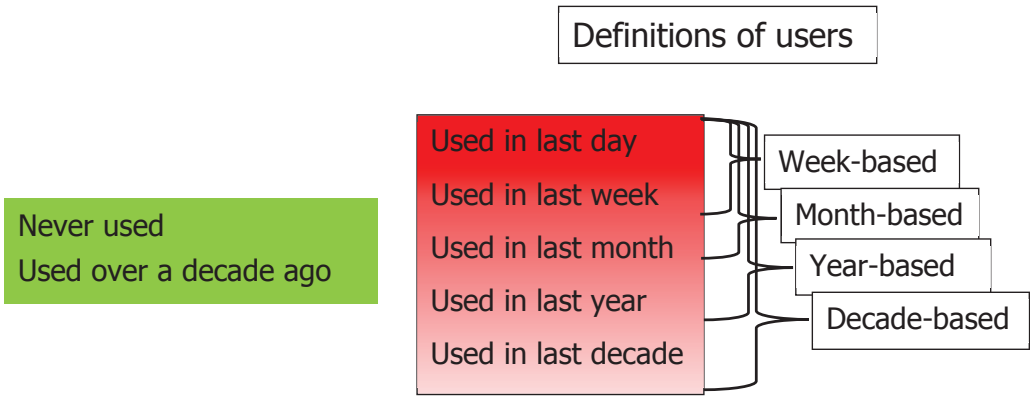

Figure 2.1. Categories of drug users. Categories with green background always correspond to drug non-users. Four different definitions of drug users are presented.

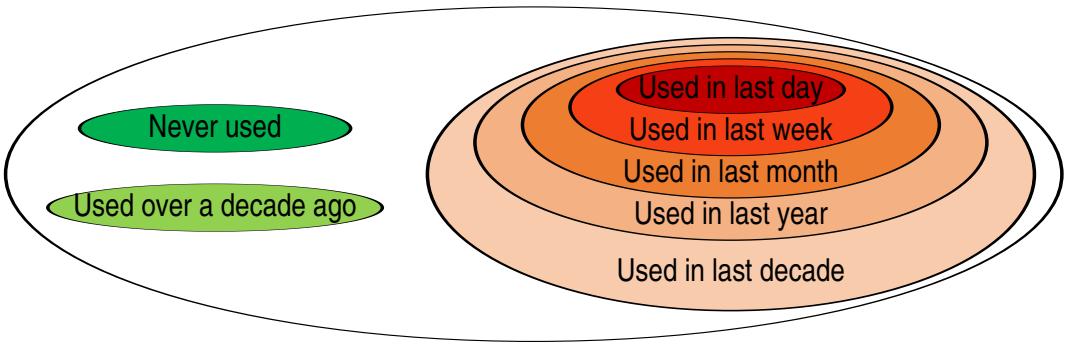

Figure 2.2. Classes of drug users.

spuriously) dichotomises individuals as users or non-users, without due regard to their frequency or duration/desistance of drug use [82]. In this study, finer distinctions concerning the measurement of drug use have been deployed, due to the potential for the existence of qualitative differences between individuals with different usage levels. In relation to each drug, respondents were asked to indicate if they never used this 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. Different categories of drug users are depicted in Fig 2.1.

Analysis of the classes of drug users shows that part of the classes are nested: participants which belong to the category 'Used in last day' also belong 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: 'Never used' and 'Used over a

decade ago' (see Fig 2.2). The data does not contain a definition of the users and non-users groups. Formally only a participant in the class 'Never used' can be called a non-user, but it is not a seminal definition because a participant who used a drug more than decade ago cannot be considered a drug user for most applications. There are several possible way to discriminate participants into groups of users and non-users for binary classification:

1. Two isolated categories ('Never used' and 'Used over a decade ago') are placed into the class of non-users with a green background in Fig 2.1, and all other categories into the class 'users' as the simplest version of binary classification. This classification problem is called '*decade-based*' user/non-user separation.

2. The categories 'Used in last decade', 'Used over a decade ago' and 'Never used' are merged to form a group of non-users and all other categories are placed into the group of users. This classification problem is called '*year-based*'.

3. The categories 'Used in last year', 'Used in last decade', 'Used over a decade ago' and 'Never used' are combined to form a group of non-users and all three other categories are placed into the group of users. This classification problem is called '*month-based*'.

4. The categories 'Used in last week' and 'Used in last day' are merged to form a group of users and all other categories are placed into the group of non-users. This classification problem is called '*week-based*'.

We begin this analysis from the decade-based user/non-user separation because it is a relatively well-balanced classification problem, that is, there are sufficiently many users in the united group 'Used in last decade-year-month-week-day' for all drugs in the database. If the problem is not directly specified then it is the decade-based classification problem. We also perform analysis for the year-,

Table 2.6. The number and fraction of drug users

Drug	User definition based on			
	Decade	Year	Month	Week
Alcohol	1817; 96.39%	1749; 92.79%	1551; 82.28%	1264; 67.06%
Amphetamines	679; 36.02%	436; 23.13%	238; 12.63%	163; 8.65%
Amyl nitrite	370; 19.63%	133; 7.06%	41; 2.18%	17; 0.90%
Benzodiazepines	769; 40.80%	535; 28.38%	299; 15.86%	179; 9.50%
Cannabis	1265; 67.11%	999; 53.00%	788; 41.80%	648; 34.38%
Chocolate	1850; 98.14%	1840; 97.61%	1786; 94.75%	1490; 79.05%
Cocaine	687; 36.45%	417; 22.12%	159; 8.44%	60; 3.18%
Caffeine	1848; 98.04%	1824; 96.76%	1764; 93.58%	1658; 87.96%
Crack	191; 10.13%	79; 4.19%	20; 1.06%	11; 0.58%
Ecstasy	751; 39.84%	517; 27.43%	240; 12.73%	84; 4.46%
Heroin	212; 11.25%	118; 6.26%	53; 2.81%	29; 1.54%
Ketamine	350; 18.57%	208; 11.03%	79; 4.19%	37; 1.96%
Legal highs	762; 40.42%	564; 29.92%	241; 12.79%	131; 6.95%
LSD	557; 29.55%	380; 20.16%	166; 8.81%	69; 3.66%
Methadone	417; 22.12%	320; 16.98%	171; 9.07%	121; 6.42%
MMushrooms	694; 36.82%	434; 23.02%	159; 8.44%	44; 2.33%
Nicotine	1264; 67.06%	1060; 56.23%	875; 46.42%	767; 40.69%
VSA	230; 12.20%	95; 5.04%	34; 1.80%	21; 1.11%

month-, and week-based user/non-user separation. It is useful to group drugs with highly correlated usage for this purpose (see Section 'Pleiades of drugs').

The proportion of drug users among all participants is different for different drugs and for different classification problems. The data set comprises 1885 individuals without any missing data. Table 2.6 shows the percentage of drug users for each drug and for each problem in the database. It is necessary to mention that the sample is intentionally biased to a higher proportion of drug users. This means that for the general population the fraction of an illegal drug users is expected to be significantly lower [83]. The standard problem of online surveys is the unintentional biasing of samples [84]. We return to this problem in the next chapter.

2.4 Data sets for AASA testing

We applied AASA for three real life datasets which have different number of records and different dimensions and are taken from different application areas:

- 'Drug consumption'(psychology, [2,3]). Detailed description of this dataset is provided in 'Database' section. Required accuracy for this database is 65%.

- 'USA president elections' (politics, [4]). This database contains 31 instances and 12 Boolean features. This dataset describes results of elections of the USA president for period from 1860 through 1980. The 12 input features are answers for the questions which concern the political, economic, social conditions of the country, and candidates themselves. List of questions is presented in Appendix D.2. Detailed description of this data set is provided in [4,169]. Required accuracy for this database is 75%.

- 'Breast cancer' (medicine, [5]. This dataset contains 569 observation of found needle aspirates of breast cancer. There are 10 measurements which computed for each nucleus: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The mean value, worst (mean of the three largest value), and standard error of each measurement are computed for each observation. As a result, there are 30 input features. Detailed description of this data set is provided in [5,61] . Required accuracy for this database is 95%.

CHAPTER 3

Materials and Methods of Data Analysis

3.1 T–scores

Transformation of all scores to T-scores is a standard normalization procedure. The raw score for each factor of the NEO-FFI-R is converted into a T-Score based on normative data [37]:

$$T\text{-}score = 10\left[\frac{Raw\ score\ -\ Normative\ mean\ score}{Normative\ standard\ deviation}\right] + 50 \qquad (3.1.1)$$

It is highlighted below that the sample in this study deviates from the population norm in the same direction as drug users deviate from sample mean (see Table 4.6). However, the deviances of mean of users groups are different for different drugs. Therefore, it is convenient to study deviations of users and non-users from the sample mean. We introduce sample based T-score for this purpose.

T-$score_{sample}$ is introduced for improving the visibility and simplicity of a comparison and is calculated using equation (3.1.2). The resulting T-$score_{sample}$ contains a mean of 50 and a standard deviation of 10.

$$T\text{-}score_{sample} = 10\left[\frac{Raw\ score\ -\ Sample\ mean\ score}{Sample\ standard\ deviation}\right] + 50 \tag{3.1.2}$$

Usually, the T-score is categorized into five categories to summarise an individual's personality score concerning each factor. The interval 20-35 indicates very low scores. The interval 35-45 indicates low scores. The interval 45-55 indicates average scores. The interval 55-65 indicates high scores. The interval 65-80 indicates very high scores. This study considers the mean T-$score_{sample}$ for two groups (users and non-users) instead of each individual's score. A subdivision of the T-$score_{sample}$ interval is introduced as follows: the interval 44-49 indicates moderately low scores $(-)$, the interval 49-51 indicates neutral scores (0), and the interval 51-56 indicates moderately high scores $(+)$.

The unification of the mean and variance of the T-$score_{sample}$ for all factors simplifies comparisons of groups (both users and non-users) for each drug. Any differences between the mean T-$score_{sample}$ for groups of users and non-users is usually used as a measure of the groups' dissimilarity in scores. The NEO-FFI-R scores for groups of users and non-users for each drug were represented by the mean T-$score_{sample}$ of these groups for each factor. A t-test is employed to estimate the significance of the differences between the mean T-$score_{sample}$ for groups of users and non-users for each NEO-FFI-R factor and each drug. In this t-test, a p-value is a probability of observing by chance the same or a greater difference of mean for two samples with the same mean. The 90% level is chosen to select significant differences. The analysis was performed using SAS 9.4.

3.2 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 many levels. To apply logistic regression to these data with categorical features and corresponding coefficients, we have to use dummy coding directly

or indirectly. In this case we have $n-1$ coefficients for a feature with n levels, meaning that we fit logistic regression in a 250 dimensional space (age contains six levels, Gndr contains two levels, education contains nine levels, country contains seven levels, ethnicity contains seven levels, N score contains 49 levels, E score contains 42 levels, O score contains 35 levels, A score contains 41 levels, C score contains 41 levels, impulsivity 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 a logistic regression model in a 12 dimensional space. This means that feature quantification can be used as an effective dimensionality reduction method.

3.2.1 Principal Component Analysis – PCA

Karl Pearson [92] invented PCA as a method for approximation of data sets by straight lines and lower-dimensional planes (Fig. 3.1). The *Unexplained Variance* is the quadratic error of PCA, that is the sum of the squared distances from the data points to the approximating low-dimensional plane or line. The *Explained Variance* is the variance of the orthogonal projections of the data on the approximating line or plane. The *Fraction of Variance Unexplained (FVU)* is the ratio of the unexplained sample variance to the total sample variance. The *Fraction of Variance Explained*, otherwise known as the *Coefficient of Determination* R^2 is 1−FVU.

PCA is equivalent to spectral decomposition of the sample covariance matrix (or to the singular value decomposition (SVD) of the data matrix). The eigenvalues λ_i of the sample covariance matrix are real and non-negative. Let us order them by size in descending order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, ($n$ is the dimension of the data space) and choose a corresponding orthonormal basis of eigenvectors: $w_1, w_2, \ldots, w_n$. These eigenvectors are called the *Principal Components* (PC for short), and the best approximation of data by a k-dimensional plane is given by the first k PCs via the formula

$$x - \bar{x} \approx \sum_{i=1}^{k} w_i(w_i, x - \bar{x}), \tag{3.2.1}$$

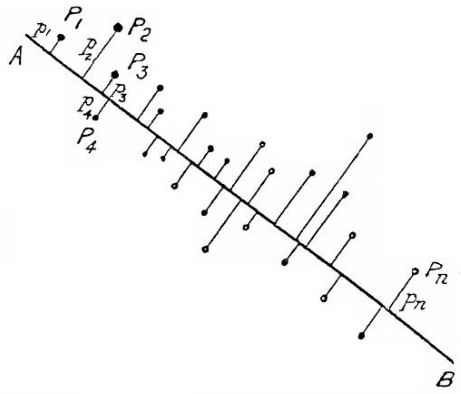

Figure 3.1. Pearson's illustration of PCA definition [92]. P_i are data points, p_i are their distances to the approximating line. The best approximation problem is UV$= \sum_i p_i^2 \to min$

where (w_i, x) is the inner product. For this approximation, FVU$= \sum_{j>k} \lambda_j$. Usually, PCA is applied after normalization and centralization to z-scores. The sample covariance matrix for $z - scores$ is the PCC (correlation) matrix.

For applications, there exists a crucial question without a definite theoretical answer: how do we select k? That is, how many PCs should be used in our approximation? There are several popular heuristic rules for component retention [93]. The most celebrated of them is Kaiser's rule: retain PCs with λ_i above the average value of λ [93,94]. The trace of a matrix A is the sum of its diagonal elements, and we denote this by tr A. We note that the average λ is $\frac{1}{n}\text{tr}\, Q = \frac{1}{n}\sum_i s_i^2$, where Q is the sample covariance matrix, and s_i^2 is the sample variance of the ith attribute. For the PCC matrix, the average λ is 1 because $r_{ii} = 1$, and Kaiser's rule is: retain PCs with $\lambda > 1$.

Various approaches to PCA were discussed in [123, 124]. There are several directions for generalisation of PCA: nonlinear PCA [91, 122, 123], branching PCA [90, 121], nonlinear PCA with categorical variables [88]. Different norms for approximation errors have been employed instead of quadratic FVU [125–128].

3.2.2 Ordinal feature quantification

One of the widely used techniques to analyse ordinal data is the calculation of polychoric correlation [86, 87]. The matrix of polychoric coefficients is used further to calculate principal components (PCs), etc. The technique of polychoric correlation is based on the assumption that values of ordinal features result from the discretization of continuous random values with fixed thresholds. Furthermore, these latent continuous random values follow a normal distribution. Unfortunately, the polychoric correlation technique has two drawbacks: it defines the thresholds of discretization but not the values for each category, and the defined thresholds differ for different pairs of attributes.

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

$$t_i = \phi^{-1}\left(\sum_{j=1}^{i} p_j\right), \tag{3.2.2}$$

where ϕ is cumulative distribution function for standard normal distribution.

The simplest method of the ordinal feature quantification is to use the thresholds (3.2.2) and select the 'average' value in each interval. There are several variants of the 'average' value. In 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 for this interval is

$$q_i = \phi^{-1}\left(\sum_{j=1}^{i-1} p_j + \frac{p_i}{2}\right) \tag{3.2.3}$$

The polychoric coefficients, calculated on base of quantification (3.2.3), have less likelihood than the polychoric coefficients calculated by using the maximum likelihood approach. The merit of this approach is the usage of the same thresholds

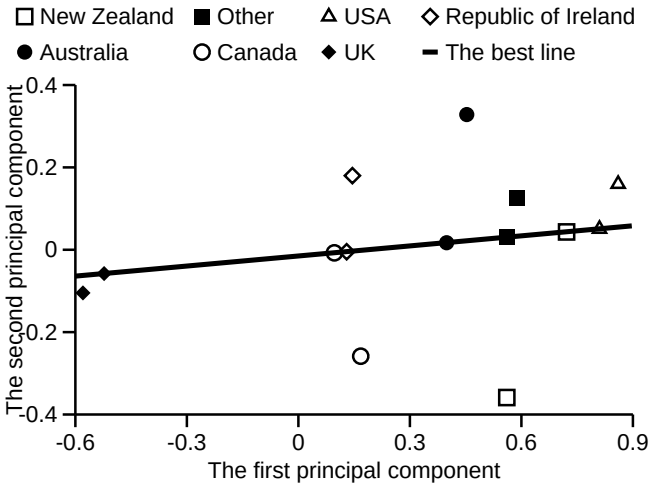

Figure 3.2. Nominal feature 'Country' quantification

for all pairs of attributes and explicit formula for calculating the categories' values.

3.2.3 Nominal feature quantification

We cannot use the techniques described above to quantify nominal features such as Gndr, country of location and ethnicity because the categories of these features are unordered. To quantify nominal features we implemented the technique of nonlinear CatPCA (Categorical Principal Component Analysis) [88]. This procedure includes four steps (Algorithm 1).

Algorithm 1 Nominal feature quantification

1: Exclude nominal features from the set of input features and calculate the informative PCs [89–92] in the space of retained input features. To select informative components we use Kaiser's rule [93, 94].
2: **for** each nominal feature **do**
3: Calculate the centroid of each category in projection on selected PCs.
4: Calculate the first PC of centroids.
5: The numerical value for each category is the projection of its centroid on this component.

The process of nominal feature quantification for the feature 'country' is depicted in Fig 3.2. Fig 3.2 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 [95] 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.

3.2.4 Input feature ranking

In this study, we used three different techniques for input feature ranking. The first technique was *principal variables* [96]. Principal variables are a set of input features which explain the maximal fraction of the data variance. The main idea of this approach is to select first the input feature which explains the maximal fraction of the data variance, then select the next feature which together with the previously selected features explains the maximal fraction of data variance, and so on.

The second technique was *double Kaiser's selection*. Calculate PCs and select informative PCs by Kaiser's rule [93, 94]. Kaiser's rule states the all PCs which correspond to eigenvalues greater than the average are informative and all other PCs are uninformative. We apply the covariance based PCs: we evaluate PCs as the normalized eigenvectors of the empirical covariance matrix. For them, the Kaiser rule threshold is equal to the trace of the covariance matrix divided by the number of attributes. The importance of an attribute is defined as the maximum of the absolute value of the corresponding coordinates in the informative PCs. In attribute selection we define the threshold of importance as the average value of coordinate which is equal to $1/\sqrt{n}$ for a unit length vector, where n is the number of attributes. If the attribute importance is greater than the threshold of importance for at least one informative PC then this attribute is informative. Otherwise, the attribute is trivial. If there are trivial attributes then the worst attribute is the attribute with minimal value of importance. We removed the worst attribute and

repeated the procedure. This procedure stops if there are no trivial attributes. This algorithm ranks attributes from worst to best.

The third technique was *sparse PCA* [97]. In this study, we used the simplest thresholding sparse PCA. The searching for each sparse PC contains several steps (Algorithm 2).

Algorithm 2 Search of sparse PC

1: Define the number of features n, variance of data σ^2, the Kaiser threshold for the components $h = \sigma^2/n$ and the Kaiser threshold for the coefficients $c = 1/\sqrt{n}$.
2: Search for the usual PC and calculate the variance σ_c^2 explained by this component. The iterative algorithm gives the PCs in descending order of σ_c^2.
3: **if** $\sigma_c^2 < h$ **then**
4: All the informative components are found. Remove the last component and go to step 10.
5: Search the attribute with non-zero coefficients with least absolute value c_{min}.
6: **if** $c_{min} > c$ **then**
7: There are no trivial attributes found. Go to step 9.
8: Set the value of the found coefficient to zero. Block changes to this attribute coefficient and search the PC under this condition. Go to step 5.
9: Subtract the projection onto the found component from the data and go to step 2.
10: Search the attributes with zero coefficients in each found component. These attributes are trivial.
11: **if** there are trivial attributes **then**
12: Remove trivial attributes from the set of attributes and go to step 1.
13: **else**
14: Stop

3.3 Methods of classification and risk evaluation

In this study, we applied several classification methods which provide risk evaluation. Each of these classification methods covers various different algorithms. We implemented this methods for such analysis with include some new idea for that. The detailed description of these classical methods is presented in this section. We provided some flowcharts for classification methods in Appendix A.

Our goal is to select the best one for the given problem. Subsequently, the best subset of input features should be chosen.

We used two general sets of input features. The first was the set of input features after quantification described in the 'Ordinal feature quantification' Section and in the 'Nominal feature quantification' Section. We called this set of feature 'original'. It includes age, Gndr, education, N, E, O, A, C, Imp, and SS. The second set was the set of projections of input features onto the first four PCs. This set of input features we called 'projected'.

3.3.1 *k* Nearest Neighbours (*k*NN)

k-Nearest neighbor algorithm (kNN) is one of the supervised machine learning algorithms that have been used in various implications in the fields of: data mining, statistical pattern recognition and some others. kNN follows a method for classifying objects based on closest training examples in the feature space.

The basic concept of *k*NN is the class of an object is the class of the majority of its k nearest neighbours [98]. k is always a positive integer. The neighbors are selected from a set of objects for which the correct classification is known. This algorithm is very sensitive to distance definition. There are several commonly used variants of distance for *k*NN: Euclidean distance; Minkovsky distance; and distances calculated after some transformation of the input space.

In this study, we used three distances: the Euclidean distance, the Fisher's transformed distance [99] and the adaptive distance [100]. Moreover, we used a weighted voting procedure with weighting of neighbours by one of the standard kernel functions [102].

Let us denote ith record of data base as $x^i = (x^i_1, \ldots, x^i_s)$, $i = 1, 2, \ldots, n$ where s is the input vector dimension, i is the number of record. Class of point x^i is $c(x^i)$. Set of all records in training is T. $y = (y_1, \ldots, y_s)$ is the test point, point class of which would be defined. k is the number of voters (nearest neighbours).

kw is the number of records, which used in distance transformation procedure. Kw is the set of records, which used in distance transformation procedure, $Kw = \{j_1, K, j_{kw}\}$.

Euclidean distance This is a classical calculation. There are no distance transformations. Euclidean distance is defined by the formula:

$$d_i = d\Big(x^i, y\Big) = \|x^i - y\| = \sqrt{\sum_{i=1}^{s}\Big(x_l^i - y_l\Big)},$$

Set of nearest neighbours is formed by the rule:

$$Kw = Sel\Big(T, y, k, d\Big)$$

where function *Sel* selects in the training set T k nearest neighbours of test point y with respect to distances d.

Fisher transformed distance The main idea of this method is the non-equivalence of directions for classification. The set of wide neighbours is defined by formulas

$$Kw = Sel\Big(T, y, kw, d\Big) \tag{3.3.1}$$

The best direction separating two classes is calculated by Fisher discriminant. The formula of the best direction is

$$\omega = \Big(\Sigma_{c=1} + \Sigma_{c=2}\Big)^{-1}\Big(\mu(Kw_1, 1) - \mu(Kw_2, 1)\Big),$$

where $\Sigma_{c=i}$ is covariation matrix of ith class wide neighbours, upper index -1 means the matrix inverse operation.

Let define the Fisher's distance as:

$$f_i = |(\omega, x^i) - (\omega, y)|$$

where (a, b) is the inner product of vectors a and b. Using this new distance we can select k nearest neighbours among wide neighbours:

$$K = Sel\Big(Kw, y, k, f\Big).$$

Adaptive distance. kNN with adaptive distance was developed by Trevor Hastie and Robert Tibshirani in 1996 [100]. The adaptive distance transformations algorithm is described in [100]. In this application we implemented a more generalized version of the proposed algorithm. In the first step the set of wide neighbours is defined by using Euclidean distance. The set of wide neighbours is selected by the rule

$$Kw = Sel\Big(T, y, kw, d\Big)$$

In the second step the matrix of distance transformation is defined. Let us consider this process in detail. Define radius of set of wide neighbours as

$$D = \max_{i \in Kw} d_i$$

For each point from set of wide neighbours define weight

$$w\big(x^i\big) = K\big(d_i/D\big)$$

where $K(x)$ is one of the statistics kernel functions [102]: uniform, triangle, Epanechnikov, quartic (biweight), tricube, triweight, Gaussian, and cosine.

Let us separate set Kw into two subsets by classes Kw_j according the formula

$$S_j = \{i : i \in S, c(x^i) = j\}, j = 1, 2$$

where S is a set, then $|S|$ number of elements in the set and $c(x^i)$ is class of point x^i.

Weighted means vector for set S and weights $w(x)$ is calculated as by formula:

$$\mu(S,w) = \frac{\sum_{i\in S} x^i w(x^i)}{\sum_{i\in S} w(x^i)}$$

and general weighted mean is also calculated by the same formula. Weighted membership of subset S in set Q is calculated by the formula

$$\pi(S,Q) = \frac{\sum_{j\in S} w(x^i)}{\sum_{j\in Q} w(x^i)}$$

Let B and W be the weighted *between*-class and *within*-class sum-of-square matrices:

$$B = \pi(Kw_1, Kw)\left(\mu(Kw_1,w) - \mu(Kw,w)\right)\left(\mu(Kw_1,w) - \mu(Kw,w)\right)^T$$

$$+\pi(Kw_2, Kw)\left(\mu(Kw_2,w) - \mu(Kw,w)\right)\left(\mu(Kw_2,w) - \mu(Kw,w)\right)^T$$

and

$$W = \frac{1}{\sum_{i\in Kw} w(x^i)}\left[\sum_{i\in Kw_1} w(x^i)\left(x^i - \mu(Kw_1,w)\right)\left(x^i - \mu(Kw_1,w)\right)^T\right.$$

$$\left. + \sum_{i\in Kw_2} w(x^i)\left(x^i - \mu(Kw_2,w)\right)\left(x^i - \mu(Kw_2,w)\right)^T\right]$$

where T is the matrix transposing symbol. New norm matrix is calculated by formula

$$N = W^{-1}BW^{-1} + \epsilon W^{-1}$$

where ϵ is the softening parameter. In this application $\epsilon = 1$ is used.

The third step of distance transformation is the calculating of the adaptive distance by the formula

$$\alpha_i = \sqrt{\left(x^i - y, N(x^i - y)\right)}$$

The last step is selecting the k nearest neighbours among instances of set Kw.

The *k*NN algorithm is well known [98]. The adaptive distance transformation algorithm is described in [100]. *k*NN with Fisher's transformed distance is less known. Algorithm of this method is presented in Algorithm 3. The following parameters are used: k is the number of nearest neighbours, K is the kernel function, and k_f is the number of neighbours which are used for the distance transformation.

Algorithm 3 *k*NN with Fisher's transformed distance

1: Find the k_f nearest neighbours of the test point.
2: Calculate the empirical covariance matrix of k_f neighbours and Fisher's discriminant direction.
3: Find the k nearest neighbours of the test point using the distance along Fisher's discriminant direction among the k_f neighbours found earlier.
4: Define the maximal distance from the test point to k neighbours.
5: Calculate the membership for each class as a sum of the points' weights. The weight of a point is the ratio: the value of the kernel function K of distance from this point to the test point divided by the maximal distance defined at step 4.
6: Drug consumption risk is defined as the ratio of the positive class membership to the sum of memberships of all classes.

The adaptive distance version implements the same algorithm but uses another transformation on step 2 and another distance on step 3 [100]. The Euclidean distance version simply defines $k_f = k$ and omits steps 2 and 3 of algorithm. We tested 1,683 million versions of the *k*NN models per drug, which differ by:

- The number of nearest neighbours, which varies between 1 and 30;
- The set of input features;
- One of the three distances: Euclidean distance, adaptive distance and Fisher's distance;
- The kernel function for adaptive distance transformation;
- The kernel functions for voting.
- Weight of class 'users' is varied between 0.01 and 5.0.

3.3.2 Decision Tree (DT)

The decision tree is one of the most popular methods of data analysis [101]. It was invented before the computer era for clarification of complex diagnosis and decision making situations, in the form of a tree of simple questions and decisions. We aim to solve the classification problem (user/non-user classification). For this purpose, we consider a decision tree as a tool for combination of various classifiers. Each elementary classifier can be thought of as a categorical variable X (symptom) with several nominal values $v_1, \ldots, v_k$. An elementary branching divides the dataset (the root) into two subsets (the nodes, see Fig. 3.3). Perfect branching creates 0-1 frequencies, for example $n_{11}/N_1 = n_{22}/N_2 = 1$ and $n_{12} = n_{21} = 0$. Such a perfect situation (errorless diagnosis by one feature) is not to be expected. However, we can find the elementary classifier, which results in the closest to the perfect classification. Then we can then iterate, i.e., approach each node as a dataset and try all possible elementary classifiers, etc., until we find a perfect solution, the solution cannot be improved, or the number of examples in a node becomes too small (which will lead to overfitting).

There are many methods for developing a decision tree [103–109], which differ depending on the method for selection of the best elementary classifier for branching (Fig. 3.3), and by the stopping criteria. We have evaluated the best attributes for branching using one of the following 'gain' functions: RIG, Gini gain, and DKM gain. They are defined in a similar way below.

Consider one node (the root, Fig. 3.3) with N cases and the binary classification problem. If the attribute X has c possible categorical values $v_1, \ldots, v_c$ then we consider branching into c nodes. We use the following notation: N_i is the number of cases with $X = v_i$ $(i = 1, \ldots, c)$, n_{ij} $(i = 1, \ldots, c,\ j = 1, 2)$ is a number of cases of class j with $X = v_i$.

For a vector of normalised frequencies $f = (f_1, \ldots, f_k)$ three concave ('Base') functions are defined:

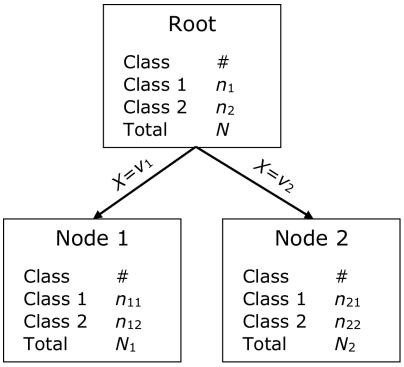

Figure 3.3. Elementary branching in a decision tree

- $Entropy(f) = -\sum_{i=1}^{k} f_i \log_2 f_i$;
- $Gini(f) = 1 - \sum_{i=1}^{k} f_i^2$;
- $DKM(f) = 2\sqrt{f_1 f_2}$ (for $k = 2$).

The corresponding gain functions for branching are defined as

$$Gain = Base\left(\frac{n_1}{N}, \frac{n_2}{N}\right) - \sum_{i=1}^{c} \frac{N_i}{N} Base\left(\frac{n_{i1}}{N_i}, \frac{n_{i2}}{N_i}\right),$$

where $Base(f)$ is one of the functions *Entropy, Gini, or DKM*.

All of these *Gain* functions qualitatively measure the same thing: how far are distributions of nodes from the initial distribution of the root, and how close they are to the perfect situation (when each node is strongly biased to one of the classes).

For a variety of reasons one might want to weight classes differently, for instance, to reduce the impact of classes with many outliers. We need to multiply class frequencies by weights and then to normalise by dividing by the sum of weights.

The branching with maximal *Gain* is considered as the best for a given criterion function.

The set of elementary classifiers (attributes) may be large, and include all the one-attribute classifiers with different thresholds, all linear discriminants, various non-linear discriminant functions, or other classifiers like kNN, etc.

The specified minimal number of instances in a tree's leaf is used as a criterion to

stop node splitting; no leaf of the tree can contain fewer than a specified number of instances.

We tested decision tree models, which differ by:

- The three split criterion (information gain, Gini gain or DKM gain);
- The use of the real-valued features in the splitting criteria separately, or in linear combination by Fisher's discriminant;
- The set of input features;
- The minimal number of instances in the leaf, which varied between 3 and 30;
- Weight of class 'users' that is varied.

There are several approaches to use real valued inputs in decision trees. A commonly used approach is the binning of real valued attributes before forming the tree. In this study we implemented 'on the fly' binning: the best threshold is searched in each node for each real valued attribute and then this threshold is used to bin these feature in this node. The best threshold depends on the split criteria used (information gain, Gini gain, or DKM gain).

Another possibility we employ is the use of Fisher's discriminant to define the best linear combination of the real valued features [99] in each node. Pruning techniques are applied to improve the tree.

The specified minimal number of instances in the tree's leaf is used as a criterion to stop node splitting. Each leaf of the tree cannot contain fewer than a specified number of instances.

Direct exhaustive search of the best decision tree can lead to the overfitiing and 'overoptimism'. Special validation procedures are necessary. The standard LOOCV meets well-known problems [117] because the topology of the tree may change in this procedure. Special notions of stability and structural stability are needed.

We return to this problem in more detail, after presentation of the classification results (Section 4.11).

3.3.3 Linear Discriminant Analysis (LDA)

The first and the most famous tool of discriminant analysis is Fisher's linear discriminant [99], where a new attribute is constructed as a linear functional of the given attributes, with the best classification ability. It is possible to calculate the score (4.1.1) for values of any function. The linear function with the highest score is a version of the Fisher's linear discriminant.

We used Fisher's linear discriminant for the binary version of the problem, to separate users from non-users of each drug. We calculate the mean of the points in the ith class, $\bar{x}_i$, and empirical covariance matrix of the ith class, S_i, for both classes ($i = 1, 2$). Then we calculate the discriminating direction as

$$\omega = (S_1 + S_2)^{-1} (\bar{x}_1 - \bar{x}_2) . \tag{3.3.2}$$

Each point x is projected onto the discriminating direction by calculating the dot product (ω, x). The threshold to separate two classes is calculated by finding the maximum of relative information gain, Gini gain, or DKM gain.

In the study we have prepared linear discriminants for all possible selections of the subset of input attributes, and selected the best set of inputs for each classification problem. we tested 8,192 LDA models per drug, which differ by one of the three criteria (information gain, Gini gain or DKM gain) which were used to define the threshold and the set of input features.

3.3.4 Gaussian Mixture (GM)

Gaussian mixture is a method of estimating probability under the assumption that each category of a target feature has a multivariate normal distribution [110].

In each category we should estimate the empirical covariance matrix and invert it. The primary probability of belonging to the ith category is:

$$p_i(x) = p_i^0 (2\pi)^{-\frac{k}{2}} |S_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x - \bar{x}_i)' S_i^{-1} (x - \bar{x}_i)\right]$$

where p_i^0 is a prior probability of the ith category, k is the dimension of the input space, $\bar{x}_i$ is the mean point of the ith category, x is the tested point, S_i is the empirical covariance matrix of the ith category and $|S_i|$ is its determinant. The final probability of belonging to the ith category is calculated as

$$p_i^f(x) = p_i(x) / \sum_j p_j(x).$$

The prior probabilities are estimated as the proportion of cases in the ith category. We also used s varied multiplier to correct priors for the binary problem.

In the study, we tested Gaussian mixture models, which differ by the set of input features and corrections applied to the prior probabilities.

3.3.5 Probability Density Function Estimation (PDFE)

We implemented the radial-basis function method [111] for probability density function estimation [112]. The number of probability densities to estimate is equal to the number of categories of the target feature. Each probability density function is estimated separately by using nonparametric techniques. The prior probabilities are estimated from the database: $p_i = n_i / N$ where n_i is the number of cases with i category of the target feature and N is the total number of cases in the database.

We also use the database to define the k nearest neighbours of each data point. These k points are used to estimate the radius of the neighbourhood of each point as a maximum of the distance from the data point to each of its k nearest neighbours. The centre of one of the kernel functions is placed at the data point [102].

The integral of any kernel function over the whole space is equal to one. The total probability of the ith category is proportional to the integral of the sum of the kernel functions, which is equal to n_i. The total probability of each category has to be equal to the prior probability p_i. Thus, the sum of the kernel functions has to be divided by n_i and multiplied by p_i. This gives the probability density estimation for each category.

We tested 426,000 versions of the PDFE models per drug, which differ by:

- The number of nearest neighbours (varied between 5 and 30);
- The set of the input features;
- The kernel function which was placed at each data points.

3.3.6 Logistic Regression (LR)

Logistic regression is a technique for analysing binary problem only. Let $X \in R^{n \times d}$ be a data matrix where n is the number of instances and d is the number of features, and y be a binary output vector $x_i \in R^d$, where $i = 1, \ldots, n$. The response variable is either the person has drug use $(y_i = 1)$ or the person has not drug use $(y_i = 0)$. The aim is to classify the instance x_i as drug use or non-use. An instance can be thought of as a Bernoulli trial with an expected value $E(y_i)$ or probability p_i. The logistic function normally used to model each instance x_i with its expected output is given as follows [113]:

$$E[y_i|x_i] = p_i = e^{(\frac{x_i\beta}{1+e^{x_i\beta}})}$$

where β is vector of parameters. We consider extended data vector (we add 1 as a first element) and include intercept into parameter vector β.

The logistic (logit) transformation is the logarithm of the odds defined as follows:

$$g(x_i) = ln\frac{p_i}{1-p_i} = x_i\beta$$

The regularized log likelihood is defined as

$$lnL(\beta) = \sum_{i=1}^{n}(y_i ln(p_i) + (1-y_i)ln(1-p_i))$$

In the sequel we have implemented the weighted version of logistic regression [113]. Recall that this method can be used for binary problems only, and is based on the following model assumption:

$$\frac{\text{probability of the first class}}{\text{probability of the second class}} = \exp(\beta, x), \tag{3.3.3}$$

where β is the vector of regression coefficients and x is a data vector.

The maximum log likelihood estimate of the regression coefficients is used. This approach assumes that the outcomes of different observations are independent and maximises the weighted sum of logarithms of their probabilities. In order to prevent class imbalance difficulties, the weights of categories are defined. The most common weight for the ith category is the inverse of the fraction of the ith category cases among all cases. Logistic regression gives only one result because there is no option to customize the method except by choice of the set of input features. We performed an exhaustive search for the best set of inputs for each classification problem.

3.3.7 Naïve Bayes (NB)

Naïve Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features. Bayes theorem provides a way of computing the class posterior probability

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y_i=1}^{C} p(x|y_i)p(y_i)}$$

The Naïve Bayes approach is based on the simple assumption: attributes are independent. Under this assumption, we can evaluate the distribution of the separate attributes and then produce a joint distribution function just as a product of attributes' distributions. Surprisingly, this approach performs satisfactorily in many real life problems despite the obvious oversimplification.

We have used the standard version of NB [114]. All attributes which contain $\leq$20 different values were interpreted as categorical and the standard contingency tables were calculated for such attributes. The calculated contingency tables are used to estimate conditional probabilities. Attributes which contain more than 20 different values were interpreted as continuous. The mean and the variance were calculated for continuous attributes instead of the contingency tables. We calculated the isolated mean and variance for each value of the output attribute. The conditional probability of a specified outcome o and a specified value of the attribute x were evaluated as the value of the probability density function for a normal distribution at point x with matched mean and variance, which were calculated for the outcome o. This method has no customization options and was tested on different sets of input features. In the study we tested 2,048 NB models per drug.

3.3.8 Random Forest (RF)

Random forests were proposed by Breiman [115] for building a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of the data [116]. The random forests classification procedure consists of a collection of tree structured classifiers $h(x, \Theta_k), k = 1, \ldots, K$, where the Θ_k are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [115].

In a random forest, each tree is constructed using a different bootstrap sample (technique for reducing the variance of an estimated prediction function) from the original data [117]. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node [118].

Before each split, it select $m \leq p$ of the input variables at random as candidates for splitting. At each tree split, a random sample of m features is drawn, and only those m features are considered for splitting. Typically $m = \sqrt{p}$ or $log_2 p$, where p is the number of features ($p = 10,4$ in our case). Random Forest does not over fit. For each tree grown on a bootstrap sample, the error rate for observations left out of the bootstrap sample is monitored. This is called the "out-of-bag" error rate OOB.

Random forests try to improve on bagging by 'de-correlating' the trees. Each tree has the same expectation [117]. The forest error rate depends on two things [115]. The first is the correlation between any two trees in the forest. Increasing the correlation increases the forest error rate. The second is the strength of each individual tree in the forest. A tree with a low error rate is a strong classifier. Increasing the strength of the individual trees decreases the forest error rate. The random forest algorithm builds hundreds of decision trees and combines them into a single model [119]. In the case study we tested 2,048 RF models per drug.

Algorithm of Random Forest for Classification as follows [117]:

1. "For b=B
    - Draw a bootstrap sample Z^* of size N from the training data.
    - Grow a random-forest tree T_b to the bootstrapped data, by re-cursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.

        i Select m variables at random from the p variables.

        ii Pick the best variable/split-point among the m.

iii Split the node into two daughter nodes.

2. Output the ensemble of trees $\{T_b\}_1^B$

   To make a prediction at a new point x:

   Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{rf}^B(x) =$ majority vote $\{\hat{C}_b(x)\}_1^b$".

3.4 Criterion for selecting the best method

A number of different criteria exist for the selection of the best classifier. The criterion we used was to pick the method such that the minimum between sensitivity and specificity was maximised. If minimum between sensitivity and specificity is the same for two classifiers, then we select the classifier with the maximal sum of the sensitivity and specificity. Classifiers with sensitivity or specificity less than 50% were not considered. Sensitivity and specificity are defined as follows [155]. Sensitivity is also referred as true positive rate or recall. It is the proportion of correctly recognised positive or true positives to total number of positives:

$$Sensitivity = \frac{number\ of\ true\ positives}{total\ number\ of\ positives}.$$

Specificity is also known as true negative rate. It is the proportion of correctly recognised negatives or true negative to total number of negatives:

$$Specificity = \frac{number\ of\ true\ negative}{total\ number\ of\ false\ positives}.$$

There are several approaches to test the quality of classifier: usage of isolated test set, n-fold cross validation and Leave-One-Out Cross Validation (LOOCV) [120]. LOOCV is used for all tests in this study. There are some problems with classifier quality estimation for the technique like decision tree and random forest. These problems are considered in details by Hastie et al [117].

AASA assumes comparison of accuracy of two models. We also used this criterion to comparison of two models for AAS. If sensitivity and specificity of AAS model are not less than required accuracy we say that this model provides required accuracy. The value of required accuracy is different for different data sets.

3.5 Visualisation of the non-linear principal components screen: Elastic maps

Elastic maps [121] provide a tool for nonlinear dimensionality reduction. By their construction, they are a system of elastic springs embedded in the data space. This system approximates a low-dimensional manifold, a model of the *principal manifold* [91, 122]. The elastic coefficients of this system allow the switch from completely unstructured k-means clustering (zero elasticity) to the estimators located closely to linear principal components (for high bending and low stretching modules). With some intermediate values of the elasticity coefficients, this system effectively approximates non-linear principal manifolds. In this section we follow [123].

Let data set be a set of vectors S in a finite-dimensional Euclidean space. The 'elastic map' is represented by a set of nodes W_j in the same space. Each datapoint $s \in S$ has a 'host node', namely the closest to s node W_j (if there are several closest nodes then one takes the node with the smallest number). The data set is divided on classes

$$K_j = \{s \mid W_j \text{ is a host of } s\}.$$

The 'approximation energy' D is the distortion

$$D = \frac{1}{2} \sum_{j=1}^{k} \sum_{s \in K_j} \|s - W_j\|^2,$$

this is the energy of the springs with unit elasticity which connect each data point with its host node.

On the set of nodes an additional structure is defined. Some pairs of nodes, (W_i, W_j), are connected by 'elastic edges'. Denote this set of pairs E. Some triplets of nodes, (W_i, W_j, W_k), form 'bending ribs'. Denote this set of triplets G.

The stretching energy is

$$U_E = \frac{1}{2}\lambda \sum_{(W_i,W_j)\in E} \|W_i - W_j\|^2,$$

The bending energy is

$$U_G = \frac{1}{2}\mu \sum_{(W_i,W_j,W_l)\in G} \|W_i - 2W_j + W_l\|^2$$

where λ and μ are the stretching and bending moduli respectively. The stretching energy is sometimes referred to as the 'membrane' term, while the bending energy is referred to as the 'thin plate' term.

For example, on the 2D rectangular grid the elastic edges are just vertical and horizontal edges (pairs of closest vertices) and the bending ribs are the vertical or horizontal triplets of consecutive (closest) vertices. The total energy of the elastic map is thus $U = D + U_E + U_G$. The position of the nodes $\{W_j\}$ is determined by the mechanical equilibrium of the elastic map, i.e. its location is such that it minimizes the total energy U.

For a given splitting of the dataset S in classes K_j minimization of the quadratic functional U is a linear problem with the sparse matrix of coefficients. Therefore, similarly to PCA or k-means, a splitting method is used:

1. For given $\{W_j\}$ find $\{K_j\}$;
2. For given $\{K_j\}$ minimize U and find $\{W_j\}$;
3. If no change then terminate. Otherwise go to step 1.

This expectation-maximization algorithm guarantees a local minimum of U. For improving the approximation various additional methods are proposed. For example, the "softening" strategy is used. This strategy starts with a rigid grids (small length, small bending and large elasticity modules λ and μ coefficients) and finishes with soft grids (small λ and μ). The training goes in several epochs, each epoch with its own grid rigidness. Another adaptive strategy is 'growing net': one starts from small amount of nodes and gradually adds new nodes. Each epoch goes with its own number of nodes.

Elastic map is a continuous manifold. It is constructed from the elastic net as its grid approximation using some interpolation procedure between nodes. For example, the simplest piecewise linear elastic map is build by triangulation and piecewise linear map. Data points are projected into the closest points of the elastic map [121].

Elastic maps and software have been applied in various areas, from bioinformatics [91, 129] to political sciences [130], financial analysis [131] and multiphase flows [132]. Examples of elastic maps for the drug consumption database are presented in the next chapter.

3.6 Alternative Attributes Set (AAS)

In many applied data mining problems (classification, regression, etc.) the set of input attributes can be reduced. The minimal set of attributes sufficient for solution of the problem may be much smaller than initial set. Nevertheless, it may be useful to use more attributes to solve the problem when some attributes from the minimal set are unavailable or have erroneous or noise contaminated values. Therefore, some alternative attributes sets (AASA) may be useful to build a good predictor.

The elementary operation of AASA is to search of minimal set of attributes which can be found by ES, FFS, and BFS. Required accuracy also has to be specified

before it. The main idea of AASA is to select minimal sets of attributes which can solve problem separately. It means that elements of one set are redundant for another set. Each such set is AAS. Then we can create two types of models (this process is illustrated in chapter 5). If there are several models which alternative each other, then the simplest way to increase robustness of classifier is usage of all alternatives as new feature set. Let us call such model *'union model'*. Another way is fitting of the several models separately and then create ensemble. We apply the same criterion of the best model selection to select the best ensemble. We consider linear ensembles only. Its outcome is defined as $\sum w_i e_i / \sum w_i$, where e_i is outcomes of ith classifiers of the first layer and w_i is its weights. Let us call such model *'ensemble model'*. For this study, we use five approaches to defined weights of ensemble models: simple voting, Nelder-Mead, Luus Jaakola, pattern search, and linear regression. These techniques are illustrated below.

The Nelder-Mead (NM)

NM is one of the well-known unconstrained optimization methods [156]. It belongs to a class of methods which do not require derivatives and which are often claimed to be robust for problems with discontinuities or where the function values are noisy [157]. However, the NM method is a heuristic search technique that can converge to non-stationary points [158] on problems that can be solved by different methods [159]. The goal of the NM algorithm is to search unconstrained minimum of the objective function $f(x)$ in space R^m. This algorithm is based on the iterative update of a simplex in R^m. A simplex X in R^m is the convex hull of $m+1$ vertices, that is, a simplex $X = \{x_i\}_{i=1,\ldots,m+1}$ is defined by its $m+1$ vertices $x_i \in R^m$ for $i = 1, 2, \ldots, m+1$. Value of function $f(x)$ is calculated in each vertex of simplex. To defined simplex NM algorithm is the iterative procedure which moves, expands, shrinks or contracts previous simplex. When all the vertices finally converge to a single point with specified accuracy ε, the stopping criterion is satisfied.

The NM algorithm uses four coefficients: reflection α , expansion β , contraction γ and shrinkage σ. When the expansion or contraction steps are performed, the shape and the size of the simplex is changed. The NM coefficients must satisfy the following inequalities [160]:

$$\alpha > 0, \beta > 1, \beta > \alpha, 0 < \gamma < 1, \ and \ 0 < \sigma < 1$$

In this study the standard values of coefficients are used:

$$\alpha = 1, \beta = 2, \gamma = 0.5, \ and \ \sigma = 0.5$$

For specified vertices $x_1, x_2, \ldots, x_{m+1}$, NM algorithm is presented in Algorithm 4.

1. **Ordering**: The vertices are sorted by increasing function values so that the best vertex has index one and the worst vertex has index $m + 1$ as follow:

$$f(x_1) \leq f(x_2) \leq \cdots \leq f(x_{m+1})$$

2. **Centroid calculation**: Define x_c as centroid of face of simplex which is opposite to the worst vertex x_{m+1}. Centroid is

$$x_c = \frac{1}{m} \sum_{i=1}^{m} x_i$$

3. **Reflection**: calculate reflection point x_r by the formula:

$$x_r = x_c + \alpha(x_c - x_{m+1})$$

   If x_r is better than the 2nd worse point but not better than the best $(f(x_1) \leq f(x_r) < f(x_m))$ then get a new simplex by replacing the worst point $x_(m + 1)$ by the reflection point x_r and go to step 1.

4. **Expanding**: If the reflection point is best point $(f(x_r) \leq f(x_1))$ then we

calculate the expansion point x_e by the formula:

$$x_e = x_c + \beta(x_r - x_c)$$

If the expansion point is better than reflection one ($f(x_e) \leq f(x_r)$) then the worst vertex (x_{m+1}) is substituted by expansion point x_e and go to step 1. Otherwise the reflection point x_r substitutes for the worst vertex (x_{m+1}) and algorithm continues from step 1.

5. **Contracting**: Now $f(x_r) \geq f(x_m)$. Algorithm perform a contraction between x_c and the better of x_{m+1} and x_r. The contraction steps are performed when the simplex is near the optimum, which allows to decrease the size of the simplex. Contraction point is

$$x_{co} = x_c + \gamma(x_{m+1} - x_c).$$

If $f(x_{co}) \leq f(x_{m+1})$ then x_{m+1} is replaced by x_{co} and go to step 1.

6. **Shrinking**: The shrinking vertices is defined by formula:

$$v_j = x_1 + \sigma(x_j - x_1); j = 2, 3, \ldots, m+1$$

New simplex is defined by vertices $x_1, v_2, v_3, \ldots, v_{m+1}$. If distances between all pairs of vertices is less than specified accuracy than algorithm stops and go to step 1 otherwise.

Luus-Jaakola (LJ)

LJ optimization is widely used direct search optimization method. The concept was introduced by Luus and Jaakola (1973) [161] . Let us consider the familiar problem of minimizing (maximizing) a real value scalar function of m variables, written as the performance index

Algorithm 4 Nelder-Mead algorithm

repeat
  Sort vertices to hold $f(x_1) \leq f(x_2) \leq \cdots \leq f(x_{m+1})$ ▷ **Ordering**
  ▷ **Centroid calculation**: Define x_c as centroid of face of simplex
  ▷ which is opposite to the worst vertex x_{m+1}.
  $x_c \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i$
  $x_r \leftarrow x_c + \alpha(x_c - x_{m+1})$ ▷ Calculate reflection point
  if $f(x_1) \leq f(x_r) < f(x_m)$ **then**
    $x(m+1) \leftarrow x_r$ ▷ **Reflection**
  else if $(f(x_r) \leq f(x_1))$ **then**
    $x_e \leftarrow x_c + \beta(x_r - x_c)$ ▷ Calculate the expansion point x_e
    if $(f(x_e) \leq f(x_r))$ **then**
      $(x_{m+1}) \leftarrow x_e$ ▷ **Expanding**
    else
      $x_m + 1 \leftarrow x_r$ ▷ **Reflection**
  else ▷ Now $f(x_r) \geq f(x_m)$
    $x_{co} \leftarrow x_c + \gamma(x_{m+1} - x_c)$. ▷ Calculate contraction point
    if $f(x_{co}) \leq f(x_{m+1})$ **then**
      $x_{m+1} \leftarrow x_{co}$ ▷ **Contracting**
    else
      $x_j \leftarrow x_1 + \sigma(x_j - x_1); j = 2, 3, \ldots, m+1$ ▷ **Shrinking**
until $\min_{i,j} ||x_i - x_j|| \geq \varepsilon$

$$I = f(x_1, x_2, \ldots, x_m) \tag{3.6.1}$$

Subject to the general inequality constraints

$$g_j(x_1, x_2, \ldots, x_m) \leq 0, j = 1, \ldots, k \tag{3.6.2}$$

and the equality constraints

$$h_i(x_1, x_2, \ldots, x_m) = 0, i = 1, \ldots, p \tag{3.6.3}$$

We assume that f, g_j, and h_i are continues functions of the variables $x_1, x_2, \ldots, x_m$. The k inequalities 3.6.1 identify the feasible region over which variables might be selected and $p < m$ in equation 3.6.3 where p is the number of equality constraints and m is the number of variables.

The LJ optimization procedure is simple, involves taking randomly chosen points over a reasonable region and then making the search more intensive around the best point by decreasing the region size in a systematic fashion (see Algorithm 5). Algorithm uses following parameters: initial point x^*, initial region-size vector r with specified radius for each coordinate, number of tested points K, a region reduction factor γ such as 0.95, required accuracy ε. The procedure is simple and FORTRAN programs using Luus Jaakola optimization procedure are given by Luus (1993, 2000) [162].

1. Select some reasonable initial point as currently optimal point x^* and reasonable initial region-size vector r^1 with specified radius for each coordinate.

2. Choose a number R of random points in the m-dimensional space around currently optimal point x^* through the equation

$$x_k^i = x_k^* + d_k^i r_k^j, i = 1, \ldots, R \tag{3.6.4}$$

   where x^* is the currently optimal point, d_k^i is chosen at random in the interval $[-1, +1]$ and r^j is the region-size vector at the jth iteration.

3. Checking of feasibility with respect to the inequality 3.6.2. All infeasible points are removed.

4. For each feasible point evaluate the performance index I in equation 3.6.1, and store the best x-value as new currently optimal point x^*.

5. Decrease the region-size vector r^j by the factor γ through

$$r^{j+1} = \gamma r^j, \tag{3.6.5}$$

Algorithm 5 Luus-Jaakola algorithm

repeat
  Randomly select d_k^i in the interval $[-1, +1]$
  $x_k^i \leftarrow x_k^* + d_k^i r_k, i = 1, \ldots, K$ ▷ Calculate test points
  Check the inequality 3.6.2. All inappropriate points are removed.
  Select the best candidate $\hat{x} = \arg\min f(x^i)$
  if $f(x^*) > f(\hat{x})$ **then**
    $x^* \leftarrow \hat{x}$ ▷ Radius reduction
  else
    $r \leftarrow \gamma r$ ▷ New centre is found
until $r \geq \varepsilon$

Pattern search (PS)

Is one of the familiar classes of methods to minimize functions without the use of derivatives or of approximations to derivatives [163, 164]. The goal of the PS algorithm is to solve the following unconstrained optimization problem:

$$\min_{x \in R^m} f(x),$$

where $f(x)$ is the objective function $f : R^m \rightarrow R$. A pattern must be form appositive spanning set for R^m [165]. In this study we apply cross pattern in R^m with $2m + 1$ points: $p_0 = (0, \ldots, 0), p_1 = (1, 0, \ldots, 0), p_{-1} = (-1, 0, \ldots, 0), \ldots, p_m = (0, \ldots, 0, 1), p_{-m} = (0, \ldots, 0, -1)$. These method have four parameters: initial central point x^*, initial pattern size r, a region reduction factor γ such as 0.95, and required accuracy ε. Pattern search algorithm is presented in Algorithm 6.

1. Calculate value of function $f(x)$ in all pattern points:

$$f_i = f(x^* + rp_i), i = (-m, \ldots, m).$$

2. Find

$$i^* = arg \min_{i=-m,\ldots,m} f_i$$

3. If $i^* = 0$ then pattern size has to be reduced: $r \longleftarrow r/2$. If r is less than

specified accuracy then algorithm stops. Otherwise go to step 1.

4. If $i^* \neq 0$ then move pattern centre to point with minimal value of function: $x^* \leftarrow x^* + rp_{i^*}$. Go to step 1.

The general form of a PS method for unconstrained minimization is described into [163, 164, 166].

Algorithm 6 Pattern search algorithm

```
repeat
    f_i ← f(x* + r p_i), i = (−m, ..., m).        ▷ Function evaluation
    i* = arg min_{i=−m,...,m} f_i                  ▷ Best candidate selection
    if f(i*) = 0 then
        r ← γ r                                    ▷ Radius reduction
    else
        x* ← x* + r p_i                            ▷ New centre is found
until r ≥ ε
```

Linear Regression

Given a data set $\{y_i, x_{i1}, \ldots, x_{im}\}_{i=1}^{n}$ where y_i response variable, x_i the m-dimensional vector of predictors, and n is the number of observations. A linear regression model assumes that the relationship between response variable y_i and the vector of predictors x_i is linear [167]. The general model takes the form

$$y_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i = x_i^T \beta + \varepsilon_i, \qquad i = 1, \ldots, n.$$

where T denotes the transpose, so that $x_i^T \beta$ is the inner product between vectors x_i and β, ε_i is error part which is independent on each predictor. This equation can be rewritten as [168].

$$y = X\beta + \varepsilon.$$

We search vector β which minimize sum of squared deviations

$$\sum_{i=1}^{n} (y_i - x_i^T \beta)^2.$$

Then we use regression coefficients as weights to form linear ensemble.

CHAPTER 4

Results of data analysis

4.1 Descriptive statistics and psychological profile of illicit drug users

The data set contains seven categories of drug users: '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'. A respondent selected their category for every drug from the list. We formed four classification problems based on the following classes (see Section 'Drug use'): the decade-, year-, month-, and week-based user/non-user separations.

We have identified the relationship between personality profiles (NEO-FFI-R) and drug consumption for the decade-, year-, month-, and week-based classification problems. We have evaluated the risk of drug consumption for each individual according to their personality profile. This evaluation was performed separately for each drug for the decade-based user/non-user separation. We have also analysed the interrelations between the individual drug consumption risks for different drugs. Part of these results has been presented in [2] (and in more detail in the 2015 technical report [2]). Section 'Pleiades of drugs' presents the notion of *correlation pleiades* of drugs. We identified three pleiades: heroin pleiad, ecstasy pleiad, and benzodiazepines pleiad, with respect to the decade-, year-, month-,

and week-based user/non-user separations. It is also important to understand how the group of users of illicit drugs differs from the group of non-users.

The descriptive statistics for seven traits (the five factor model, Imp, and SS) are presented in Table 4.1: means, standard deviations, and 95% confidence intervals for means for NEO-FFI-R for the full sample and for two subsamples: non-users of illicit drugs and users of illicit drugs. We call here 'illicit' the following drugs: amphetamines, amyl nitrite, benzodiazepines, cannabis, cocaine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, MMushrooms, and VSA. This is certainly abuse of language because in some regions the recreation use of some of these substances is decriminalized and in some countries alcohol, for example, is illicit. It may be better to call use of these drugs 'socially condemned' but this term is also not exactly defined. We use the term 'illicit drugs' just for short, while the exact definition of this group is just the list above.

A dimensionless variable z is convenient for representing the difference between two groups S_1 and S_2, with means μ_1, μ_2 and standard deviation σ_1, σ_2 respectively:

$$z = \frac{|\mu_1 - \mu_2|}{\sigma_1 + \sigma_2}. \tag{4.1.1}$$

This score measures how the variability between group means relates to the variability of the observations within the groups. It is a useful measure of the separation of the two distributions [133]. It is presented in the tables below for each of seven psychological attributes and for two classes, non-users and users of illicit drugs. The higher the score is, the better users are separated from non-users by the values of the attribute. It is possible to calculate this score for values of any function. The linear function with the highest score is a version of the Fisher linear discriminant [134]. The score z, its reciprocal z^{-1} and their multidimensional generalizations [133] are also widely used in cluster analysis for the construction of criteria for the validity of clusters [49].

A simple motivation for this measure is that for the normal distributions inside groups the optimally balanced separation of groups with given z has equal speci-

Table 4.1. Descriptive statistics: Means, 95% CIs for means, and standard deviations for the whole sample, for non-users of illicit drugs and for users of illicit drugs. Dimensionless score z (4.1.1) for separation of users from non-users of illicit drugs is presented as well as the sensitivity and specificity P of the best separation of normal distributions with this z. Sensitivity (Sn) and Specificity (Sp) are calculated for all one-feature classifiers. Θ is the threshold for class separation: one class is given by the inequality score$\leq \Theta$ and another class by score$> \Theta$.

Factors	Total sample			Non-users of illicit drugs			Users of illicit drugs			One feature classifier				
	Mean	95% CI	SD	Mean	95% CI	SD	Mean	95% CI	SD	z	P(%)	Θ	Sn (%)	Sp (%)
Decade-based separation														
N	23.92	23.51, 24.33	9.14	21.00	20.29, 21.71	7.85	24.88	24.40, 25.37	9.32	0.226	59	22	60	58
E	27.58	27.27, 27.88	6.77	28.52	27.99, 29.04	5.73	27.27	26.90, 27.63	7.05	0.098	54	28	51	55
O	33.76	33.47, 34.06	6.58	30.22	29.67, 30.77	6.06	34.93	34.60, 35.26	6.32	0.381	65	32	63	67
A	30.87	30.58, 31.16	6.44	32.87	32.35, 33.38	5.65	30.21	29.87, 30.55	6.54	0.218	59	31	61	56
C	29.44	29.12, 29.75	6.97	32.89	32.39, 33.40	5.55	28.30	27.93, 28.66	7.01	0.366	64	31	65	65
Imp	3.80	3.70, 3.90	2.12	2.74	2.58, 2.90	1.74	4.15	4.04, 4.26	2.12	0.364	64	3	71	56
SS	5.56	5.44, 5.68	2.70	3.77	3.56, 3.98	2.32	6.15	6.02, 6.28	2.55	0.490	69	4	63	74
Year-based separation														
N	23.92	23.51, 24.33	9.14	22.26	21.66, 22.87	8.15	25.18	24.64, 25.73	9.51	0.165	57	35	60	56
E	27.58	27.27, 27.88	6.77	28.21	27.76, 28.65	6.01	27.17	26.76, 27.58	7.15	0.078	53	40	51	55
O	33.76	33.47, 34.06	6.58	30.89	30.44, 31.34	6.11	35.63	35.28, 35.98	6.14	0.387	65	45	63	67
A	30.87	30.58, 31.16	6.44	32.15	31.73, 32.57	5.71	29.96	29.58, 30.33	6.58	0.178	57	43	61	58
C	29.44	29.12, 29.75	6.97	31.96	31.51, 32.42	6.16	27.77	27.37, 28.18	7.03	0.318	62	42	65	63
Imp	3.80	3.70, 3.90	2.12	2.97	2.84, 3.11	1.84	4.32	4.20, 4.44	2.11	0.342	63	3	71	59
SS	5.56	5.44, 5.68	2.70	4.00	3.82, 4.19	2.46	6.50	6.36, 6.63	2.40	0.514	70	5	63	69
Month-based separation														
N	23.92	23.51, 24.33	9.14	22.60	22.06, 23.14	8.22	25.13	24.53, 25.73	9.69	0.141	56	35	56	60
E	27.58	27.27, 27.88	6.77	28.27	27.89, 28.65	5.78	27.19	26.74, 27.64	7.27	0.083	53	39	56	51
O	33.76	33.47, 34.06	6.58	31.01	30.60, 31.43	6.24	36.04	35.67, 36.41	6.00	0.411	66	45	63	63
A	30.87	30.58, 31.16	6.44	31.82	31.43, 32.22	6.02	29.81	29.40, 30.22	6.65	0.159	56	43	56	61
C	29.44	29.12, 29.75	6.97	31.78	31.38, 32.17	5.97	27.50	27.06, 27.95	7.15	0.326	63	42	61	65
Imp	3.80	3.70, 3.90	2.12	3.04	2.91, 3.16	1.92	4.43	4.30, 4.56	2.08	0.348	64	3	64	71
SS	5.56	5.44, 5.68	2.70	4.15	3.99, 4.32	2.51	6.68	6.54, 6.83	2.35	0.521	70	5	67	63
Week-based separation														
N	23.92	23.51, 24.33	9.14	22.79	22.27, 23.31	8.47	25.21	24.56, 25.86	9.72	0.133	55	35	55	60
E	27.58	27.27, 27.88	6.77	28.34	27.97, 28.71	5.96	27.09	26.60, 27.58	7.33	0.094	54	39	56	51
O	33.76	33.47, 34.06	6.58	31.18	30.78, 31.57	6.37	36.17	35.77, 36.57	5.99	0.404	66	46	66	63
A	30.87	30.58, 31.16	6.44	31.63	31.25, 32.01	6.18	29.80	29.35, 30.24	6.67	0.143	56	42	60	61
C	29.44	29.12, 29.75	6.97	31.57	31.20, 31.95	6.10	27.41	26.94, 27.89	7.10	0.315	62	41	64	65
Imp	3.80	3.70, 3.90	2.12	3.10	2.99, 3.22	1.94	4.45	4.31, 4.59	2.08	0.335	63	3	61	71
SS	5.56	5.44, 5.68	2.70	4.27	4.12, 4.43	2.52	6.73	6.58, 6.89	2.35	0.505	69	5	62	63

ficity and sensitivity P, where $P = \phi(z)$ and $\phi(z)$ is the cumulative distribution function of the standard normal distribution (with $\mu = 0, \sigma = 1$).

Let us consider decade-based separation. The first result is: users of illicit drugs differ from non-users across all seven scales. The 95% CI for means in these groups do not intersect - not sure about this - the confidence intervals for E intersect. The most significant difference was found for SS, then for O, for Imp and C, and for N. The smallest difference was found for E. Later we will demonstrate that E for users of different drugs may deviate from E for non-users in a number of different ways. Moreover, the 95% CIs for means of all three groups, total

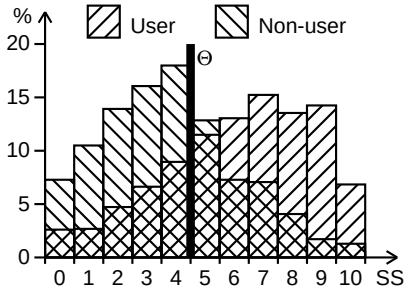

Figure 4.1. The distributions of SS for users and non-users of illicit drugs (normalized to 100% in each group) for the decade-based user/non-user separation. The optimal threshold is $\Theta = 4$

.

sample, users of illicit drugs and non-users of illicit drugs do not intersect for N, O, A, C, Imp and SS. Table 4.1 allows us to claim that the profile of users of illicit drugs has a characteristic form:

$$\mathrm{N} \Uparrow, \mathrm{O} \Uparrow, \mathrm{A} \Downarrow, \mathrm{C} \Downarrow, \mathrm{Imp} \Uparrow, \mathrm{SS} \Uparrow . \tag{4.1.2}$$

The *P* column in Table 4.1 gives a simple estimate of the separability users from non-users of illicit drugs by a single trait. The best separation is given by the value of SS: estimated sensitivity and specificity are 69%. *According to this estimate, SS for 69% of illicit drug users is higher then SS of 69% of non-users*. Of course, for more precise estimation methods the numbers will differ. We might also expect that the use of several attributes and more sophisticated classification approaches would give better sensitivity and specificity. Nevertheless, the *P* column gives us a good indication of the possible performance of classification.

Separation of classes "users of illicit drugs" and "non-users of illicit drugs" are presented in Table 4.1, where Θ is the *threshold* of the separation: one class is given by the inequality score$\leq \Theta$ and another class by score$> \Theta$. This convention, where to use strong inequality, is important because the values of scores are integer. The histogram for separation by values of SS for the decade-based definition of users is presented in Fig. 4.1.

Table 4.2 gives *p* values, i.e. the probabilities of finding the same or larger differ-

Table 4.2. Significance of differences of means for total sample, users and nonusers of illicit drugs (p-values).

Factors	Total/User	Total/Nonuser	User/Nonuser
Decade-based separation			
N	0.003	< 0.001	< 0.001
E	0.204	0.002	< 0.001
O	< 0.001	< 0.001	< 0.001
A	0.004	< 0.001	< 0.001
C	< 0.001	< 0.001	< 0.001
Imp	< 0.001	< 0.001	< 0.001
SS	< 0.001	< 0.001	< 0.001
Year-based separation			
N	< 0.001	< 0.001	< 0.001
E	0.122	0.049	0.003
O	< 0.001	< 0.001	< 0.001
A	< 0.001	< 0.001	< 0.001
C	< 0.001	< 0.001	< 0.001
Imp	< 0.001	< 0.001	< 0.001
SS	< 0.001	< 0.001	< 0.001
Month-based separation			
N	0.001	0.002	< 0.001
E	0.166	0.024	0.002
O	< 0.001	< 0.001	< 0.001
A	< 0.001	0.003	< 0.001
C	< 0.001	< 0.001	< 0.001
Imp	< 0.001	< 0.001	< 0.001
SS	< 0.001	< 0.001	< 0.001
Week-based separation			
N	0.001	0.011	< 0.001
E	0.100	0.016	0.001
O	< 0.001	< 0.001	< 0.001
A	< 0.001	0.018	< 0.001
C	< 0.001	< 0.001	< 0.001
Imp	< 0.001	< 0.001	< 0.001
SS	< 0.001	< 0.001	< 0.001

ence between mean values of the traits between users of illicit drugs, non-users of illicit drugs, and the total sample. This table completely supports the profile from (4.1.2).

There are both similarity and an important qualitative difference from the *'dark triad' of personality*, Machiavellianism, Narcissism and Psychopathy [73]. According to table 2.4, the dark triad is associated with N$\Uparrow$, A$\Downarrow$ and C$\Downarrow$ and *low* (or

Table 4.3. PCC for NEO-FFI-R for raw data

Factors	N	E	O	A	C
N		-0.432*	0.017	-0.215*	-0.398*
E	-0.432*		0.236*	0.159*	0.318*
O	0.017	0.236*		0.033	-0.060**
A	-0.215*	0.159*	0.033		0.249*
C	-0.398*	0.318*	-0.060**	0.249*	

*$p < 0.001$; **$p < 0.01$.

Table 4.4. Polychoric correlation coefficients (PoCC) of measured psychological traits (n=1885).

	N	E	O	A	C	Imp	SS
N		−0.431*	0.010	−0.217*	−0.391*	0.174*	0.080**
E	−0.431*		0.245*	0.157*	0.308*	0.114*	0.210*
O	0.010	0.245*		0.039	−0.057***	0.278*	0.422*
A	−0.217*	0.157*	0.039		0.247*	−0.230*	−0.208*
C	−0.391*	0.308*	−0.057***	0.247*		−0.335*	−0.229*
Imp	0.174*	0.114*	0.278*	−0.230*	−0.335*		0.623*
SS	0.080**	0.210*	0.422*	−0.208*	−0.229*	0.623*	

*$p < 0.0001$, **$p < 0.001$, ***$p < 0.02$.

neutral) O, whereas the profile (4.1.2) of the users of illicit drugs has the same N⇑, A⇓ and C⇓ but *high* O⇑.

Pearson's correlation coefficient (PCC or r) is employed as a measure of the strength of a linear association between two factors. PCC for all pairs of factors are presented in Table 4.3. Two pairs of factors do not have significant correlation: (1) N and O (r=0.017, p=0.471); (2) A and O (r=0.033, p=0.155). However, all other pairs of personality factors are significantly correlated in the sample (compare to Table 2.3).

Strictly speaking, the scores should be considered as ordinal features. Therefore, the polychoric correlation coefficients (PoCC) should be used. Table 4.4 presents values of PoCCs between all of the psychological traits we measured (compare to Tables 2.3 and 4.3).

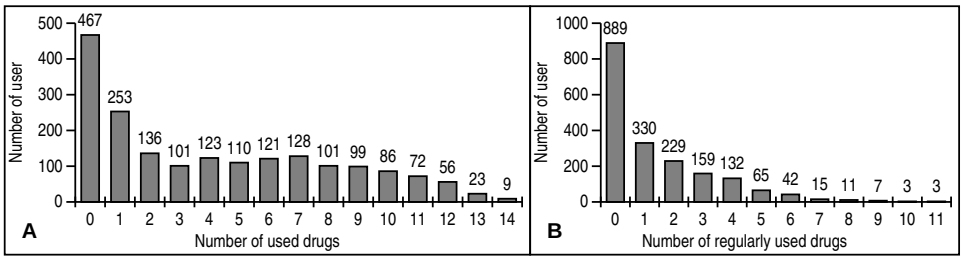

Figure 4.2. The histograms of the number of users: A - for the decade-based user/non-user separation, B - for the month-based user/non-user separation

4.2 Distribution of number of drugs used

The diagrams in Fig 4.2 show the graph of the number of users versus the number of illegal drugs used for the decade-based (A) and month-based (B) user/non-user separations. In Fig 4.2 A we can see that the distribution of the number of users is bimodal with maxima at zero and 7 drugs. In Fig 4.2 B the distribution of the number of regular users of illegal drugs looks like the exponential distribution.

The distributions of the number of users for each drug are presented in Fig 4.3 and Fig 4.4. Most of the distributions have an exponential-like shape, but several have bimodal distributions. The distributions of the number of users for the three legal drugs have maximum at 'Used in last day' or 'Used in last week' (see Fig 4.3 A, E, and G). The distribution of the number of nicotine users (smokers) has three maxima: 'Used in last day' for smokers, 'Used in last decade' for smokers who have quit smoking, and 'Never used' (see Fig 4.4 G). All distributions for illegal drug users have maximum in the category 'Never used'. However, the distribution of cannabis users has two maxima. The main maximum is in the category 'Used in last day', and the second is in the category 'Never used' (see Fig 4.3F).

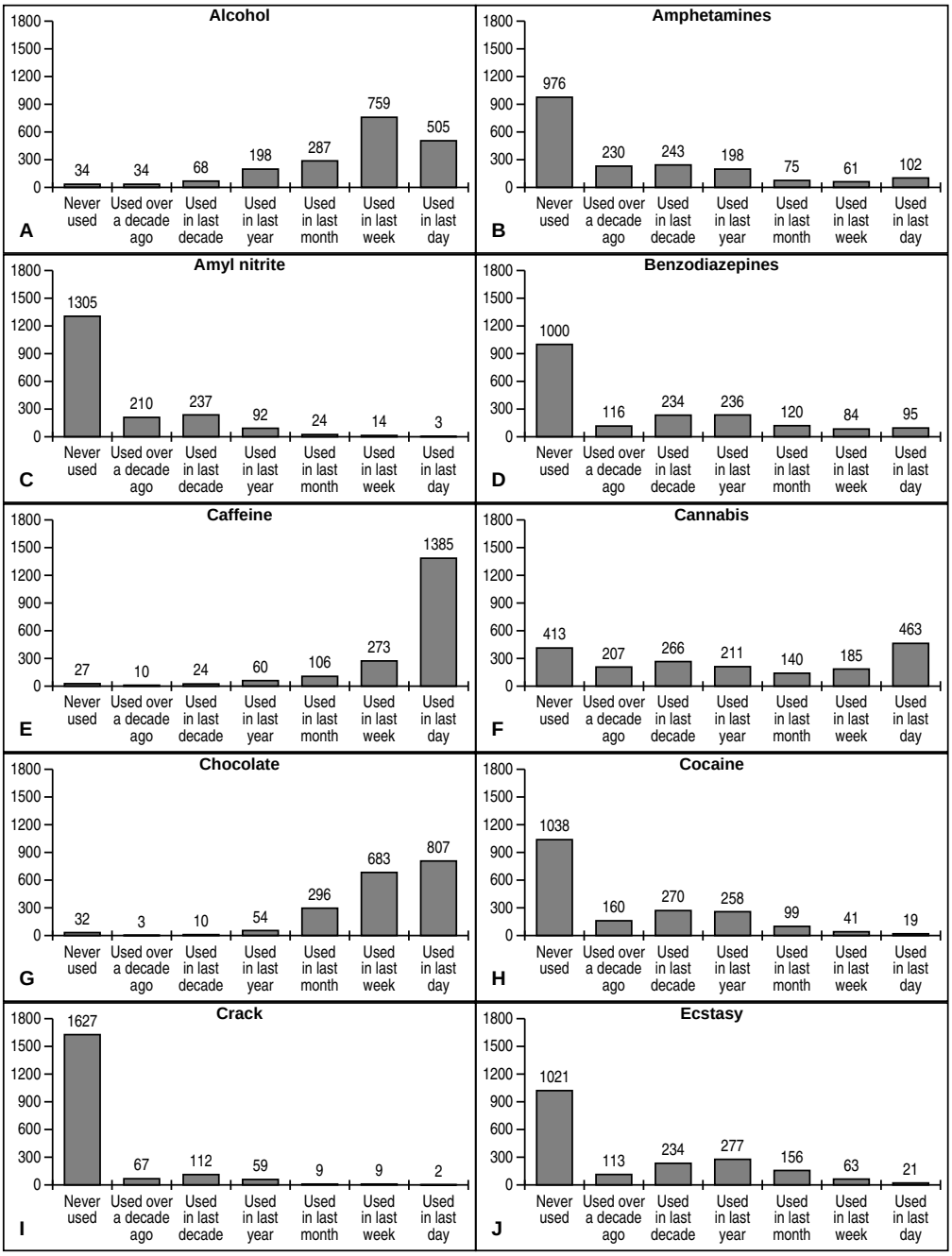

Figure 4.3. Distribution of drug usage: A: Alcohol, B: Amphetamines, C: Amyl nitrite, D: Benzodiazepines, E: Cannabis, F: Chocolate, G: Cocaine, H: Caffeine, I: Crack, and J: Ecstasy

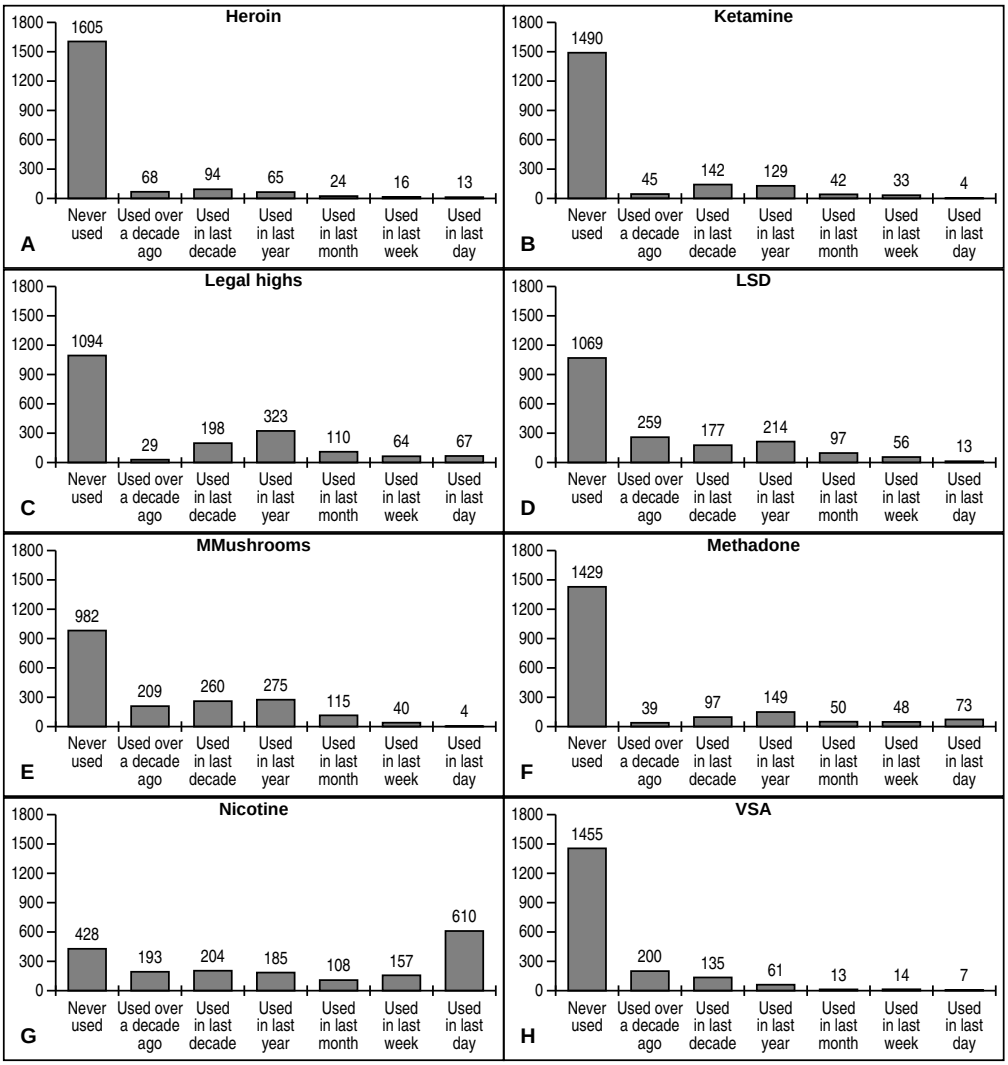

Figure 4.4. Distribution of drug usage: A: Heroin, B: Ketamine, C: Legal highs, D: LSD, E: Methadone, F: Magic mushrooms, G: Nicotine, and H: VSA

4.3 Sample mean and population norm

It may seem to be a good idea to use the T-scores with respect to the population norm. Caution is needed however, since the mean values may depend on age and social group, so that the notion of 'population norm' is a complex hierarchical construct rather than a simple set of means.

Following [37] we include data about two groups. The first consists of high school students (n=1959) [135]. The age range is from 14 to 18 (M=16.5, S.D.=1.0 years); approximately two-thirds were girls. In Table 4.5, we denote this group with the abbreviation HS.

The second sample consists of adults from the Baltimore Longitudinal Study of Aging (BLSA) [136]. BLSA participants are generally healthy and well-educated men and women who have volunteered to return to the Gerontology Research Center for periodic medical and psychosocial testing. The data was collected between 1991 and 2002, n=1492 (695 men and 797 women) aged 19–93 (M=56.2, S.D.=17.0 years); 65.1% of the sample was white, 27.6% black, and 7.3% other race. In Table 4.5 we refer to this group as BLSA.

The third sample includes 1025 participants (803 males and 221 females) and combines data from several studies published between 1996–2000 [41]. This cohort had a good range of skills, mental ability, and psychopathology, and aimed to be a representative cross-section of British society. In Table 4.5, we use Brit to label this group.

For HS and BLSA data both NEO-FFI and NEO-FFI-R profiles are available. For Brit only the NEO-FFI scores are available. For ease of comparison we include the five factor scores for the sample and for the subsample of illicit drug users (Table 4.1) in Table 4.5.

The means of the NEO-FFI-R T-scores based on normative data are depicted in Fig 4.5. For this example, the 'norm' is taken from the BLSA group (NEO-FFI). It is obvious from this figure that sample mean is significantly biased when com-

Table 4.5. Mean values of five factors for the three 'normal' samples and for the data. N-u, Illicit stands for non-users of illicit drugs with decade-based definition of users (they either never used illicit drugs or used them more than a decade ago), Samp stands for the total sample, U Illicit stands for users of illicit drugs for decade-based definition of users; compare with Table 4.1

		N		E		O		A		C	
Group	Version	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
BLSA	NEO-FFI	15.77	7.47	28.50	6.26	29.32	6.11	33.39	4.98	33.48	6.36
	NEO-FFI-R	16.83	7.36	29.29	6.46	31.29	6.12	32.41	5.42	33.26	6.30
HS	NEO-FFI	24.65	8.07	30.58	6.67	28.40	6.57	28.31	6.34	27.45	7.30
	NEO-FFI-R	25.08	7.95	31.80	6.94	31.18	6.96	28.09	6.93	27.00	7.40
Brit	NEO-FFI	19.5	8.6	27.1	5.9	26.5	6.5	29.7	5.9	32.1	6.6
N-u, Illicit	NEO-FFI-R	21.00	7.85	28.52	5.73	30.22	6.06	32.87	5.65	32.89	5.55
Samp	NEO-FFI-R	23.92	9.14	27.58	6.77	33.76	6.58	30.87	6.44	29.44	6.97
U, Illicit	NEO-FFI-R	24.88	9.32	27.27	7.05	34.93	6.32	30.21	6.54	28.30	7.01

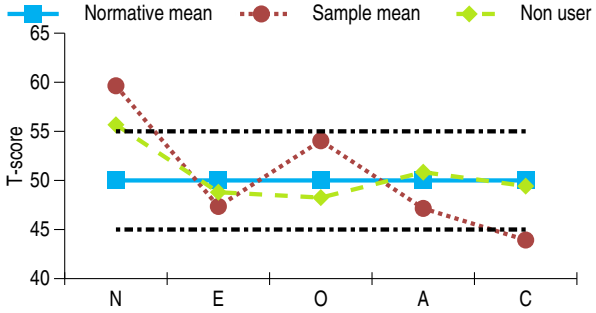

Figure 4.5. Mean T-score NEO-FFI-R for the total sample and for non-users of illicit drugs with respect to the BLSA mean as a norm.

pared to the population (represented by the BLSA group). Such a bias is usual for clinical cohorts, for example, the 'problematic' or 'pathological' groups [85], and the drug users [24,26]. For the group of non-users of illicit drugs the bias is much smaller.

It is important to observe that the mean values of scores for four factor in this sample, N, E, A, and C, are *between* the mean scores for BLSA and HS. The mean O score in this sample is significantly *higher*.

The special role of "openness to experience" (O) for drug consumption has been observed by many researchers. For example, in the paper "Undergraduate marijuana and drug use as related to openness to experience" [137] we read: "Marijuana use, in the present sample of college students, was associated with personality characteristics which many would tend to value (e.g., creativity). Open mindedness or 'openness to experience' may account for our findings. People

who are open to new experience become creative, try marijuana, and, in general, experience more than people who have a less open life style. This accounts for our finding that the more a person uses marijuana, the more likely they are to try one or more other drugs." Non-users were characterised as "the typical non-creative, high authoritarian individuals."

The high variability of means in the 'normal' groups has encouraged us to analyse the T-scores with respect to the sample mean and to study the differences between users and non-users rather than deviation from the norm. This analysis is presented in the next two sections.

4.4 Deviation of the groups of drug users from the sample mean

Tables C.1, C.2, C.3, and C.4 demonstrate the mean T-$score_{sample}$ of five NEO-FFI-R factors, supplemented by Imp and SS for users and non-users, for each drug with respect to the decade-, year-, month-, and week-based classification problems respectively (see Appendix C). Significant differences in personality factor scores are observed between these groups. The hypothesis about the universal relationship between personality profile and the risk of drug consumption can generally be described as in (4.1.2): an increase in scores of N, O, Imp, and SS suggest an increase in the risk of use, whereas an increase in the scores of A and C results in a decrease in the risk of use. Thus for each drug, drug users scored higher on N and O, and lower on A and C, when compared to non-users of drugs. The influence of the score of E is drug specific (non-universal).

We now analyse the sign of T-$score_{sample}$ for various drugs and the definitions of users (decade-, year-, month- and week-based). We used a + sign for moderately high and high T-$score_{sample}$ (T-$score_{sample} > 51$), and a – sign for moderately low and low T-$score_{sample}$ (T-$score_{sample} < 49$), and 0 for a score close to mean value $51 \geq$T-$score_{sample} \geq 49$. These signs reflect the distance from the group of users

to the sample mean.

In the next section we analyse significance of deviations. There is a standard and well-known problem with reporting of the values and significance of deviations: 'significant' does not mean 'large' and, reverse, the apparently large deviations could be insignificant. Everything is defined by interplay between the number of elements in classes and the deviation value. Practically, significant but small deviations may be unstable: after small change of conditions they may even change their sign. We can have a look on the variability of the 'normal' scores in Section 4.3. Insignificant but apparently large deviation may become significant for larger samples or may vanish. Therefore, it is necessary to answer two questions: how large is the deviation (this Section) and how significant is it (next Section)?

For a significant deviation of users from non-users we use the signs $\Uparrow$ and $\Downarrow$.

The inclusion of moderate subcategories of T-$score_{sample}$ as suggested above enables us to separate the drugs into five groups for the decade-based user/non-user separation. These are presented in Table 4.6. Each group can be coded using the (N, E, O, A, C, Imp, SS) profile:

- The group with the profile $(0, 0, 0, 0, 0, 0, 0)$ includes the users of three legal drugs: alcohol, chocolate and caffeine.
- The group of drugs with the profile $(0, 0, +, -, -, +, +)$ includes the users of amyl nitrite, LSD, and magic mushrooms.
- Nicotine users form their own group with the profile $(0, 0, +, 0, -, +, +)$.
- The largest group of drugs with the profile $(+, 0, +, -, -, +, +)$ includes the users of amphetamines, benzodiazepines, cannabis, cocaine, ecstasy, ketamine and legal highs.
- Finally, the group with the profile $(+, -, +, -, -, +, +)$ includes the users of crack, heroin, VSA and methadone.

Table 4.6. Deviation of T-$score_{sample}$ from the sample mean for various groups of users for the decade-based user/non-user separation

Drug	N	E	O	A	C	Imp	SS
Alcohol, Chocolate, Caffeine	0	0	0	0	0	0	0
Amyl nitrite, LSD, and Magic Mushrooms	0	0	+	−	−	+	+
Nicotine	0	0	+	0	−	+	+
Amphetamines, Benzodiazepines, Cannabis, Cocaine, Ecstasy, Ketamine, and Legal highs	+	0	+	−	−	+	+
Crack, Heroin, VSA, and Methadone	+	−	+	−	−	+	+
Ecstasy pleiad	0	0	+	−	−	+	+
Heroine pleiad, Benzodiazepines pleiad	+	0	+	−	−	+	+
Illicit drugs	+	0	+	−	−	+	+

For the year-based user/non-user classification, drugs are separated into eight groups as presented in Table 4.7. Each group can be coded using the (N, E, O, A, C, Imp, SS) profile:

- The group with the profile $(0, 0, 0, 0, 0, 0, 0)$ includes the users of three legal drugs alcohol, chocolate and caffeine.
- The group of drugs with the profile $(0, 0, +, -, -, +, +)$ contains just the users of magic mushrooms.
- The LSD users also form their own group with the profile $(0, 0, +, 0, -, +, +)$.
- The group with the profile $(+, 0, +, -, -, +, +)$ includes the users of amphetamines, amyl nitrite, cannabis, cocaine, crack, legal highs and VSA.
- The group of drugs with the profile $(+, -, +, -, -)$ includes the users of benzodiazepines, heroin, and methadone.
- The ecstasy users form their own group with the profile $(0, +, +, -, -, +, +)$.
- The ketamine users form their own group with the profile $(+, +, +, -, -, +, +)$.
- The nicotine users form their own group with the profile $(+, 0, +, 0, -, +, +)$.

Similarly, the deviations of T-$score_{sample}$ from the sample mean for the month-based user/non-user classification and for the week-based user/non-user classification are described in Table 4.8 and Table 4.9 respectively.

Table 4.7. Deviation of T-$score_{sample}$ from the sample mean for various groups of users for the year-based user/non-user separation

Drug	N	E	O	A	C	Imp	SS
Alcohol, Chocolate, Caffeine	0	0	0	0	0	0	0
Magic Mushrooms	0	0	+	−	−	+	+
LSD	0	0	+	0	−	+	+
Amphetamines, Amyl nitrite, Cannabis, Cocaine, Crack, Legal highs, and VSA	+	0	+	−	−	+	+
Benzodiazepines, Heroin, and Methadone	+	−	+	−	−	+	+
Ecstasy	0	+	+	−	−	+	+
Ketamine	+	+	+	−	−	+	+
Nicotine	+	0	+	0	−	+	+
Heroine pleiad, Ecstasy pleiad, Benzodiazepines pleiad	+	0	+	−	−	+	+
Illicit drugs	+	0	+	−	−	+	+

Table 4.8. Deviation of T-$score_{sample}$ from the sample mean for various groups of users for the month-based user/non-user separation

Drug	N	E	O	A	C	Imp	SS
Alcohol, Chocolate, Caffeine	0	0	0	0	0	0	0
Cannabis and Magic Mushrooms	0	0	+	−	−	+	+
Nicotine	+	0	+	0	−	+	+
Amphetamines, Ketamine, and Legal highs	+	0	+	−	−	+	+
Benzodiazepines, Heroin, and Methadone	+	−	+	−	−	+	+
Ecstasy and LSD	0	+	+	−	−	+	+
Cocaine and VSA	+	+	+	−	−	+	+
Amyl nitrite	0	0	0	−	−	+	+
Crack	+	−	0	−	−	+	+
Ecstasy pleiad	0	0	+	−	−	+	+
Heroine pleiad, Benzodiazepines pleiad	+	−	+	−	−	+	+
Illicit drugs	+	0	+	−	−	+	+

The personality profiles are strongly associated with membership of groups of the users and non-users of the 18 drugs. We found that the N and O score of drug users of all 18 drugs are moderately high $(+)$ or neutral (0), and that the A and C scores of drug users are moderately low $(-)$ or neutral (0). Detailed results can be seen in Tables 4.6, 4.7, 4.8, and 4.9.

The effect of the E score is drug specific. Drugs are divided into three groups with respect to the E score of users (in the year-, month-, and week-based classification problems) (see Tables 4.6, 4.7, 4.8 and 4.9). For example, for the week-based user/non-user separation the E score is:

Table 4.9. Deviation of T-$score_{sample}$ from the sample mean for various groups of users for the week-based user/non-user separation

Drug	N	E	O	A	C	Imp	SS
Alcohol, Chocolate, Caffeine	0	0	0	0	0	0	0
Cannabis	0	0	+	−	−	+	+
LSD and Magic Mushrooms	0	+	+	0	−	+	+
Ketamine	0	−	+	−	−	+	+
Amphetamines, Benzodiazepines, Heroin, Legal highs, and Methadone	+	−	+	−	−	+	+
Ecstasy	0	+	+	−	−	+	+
VSA	0	+	+	−	0	+	+
Cocaine	+	+	+	−	−	+	+
Nicotine	+	0	+	0	−	+	+
Amyl nitrite	0	−	0	−	−	+	+
Crack	+	−	−	−	−	+	+
Heroine pleiad, Benzodiazepines pleiad	+	−	+	−	−	+	+
Ecstasy pleiad	0	0	+	−	−	+	+
Illicit drugs	+	0	+	−	−	+	+

- Moderately low $(-)$ in groups of users of amphetamines, amyl nitrite, benzodiazepines, heroin, ketamine, legal highs, methadone, and crack;
- Moderately high $(+)$ in groups of users of cocaine, ecstasy, LSD, magic mushrooms, and VSA;
- Neutral (0) in groups of users of alcohol, caffeine, chocolate, cannabis, and nicotine.

4.5 Significant differences between groups of drug users and non-users

Tables 4.10, 4.11, 4.12, and 4.13 show where there are significant differences between the means of the personality traits for the groups of users and non-users for the decade-, year-, month-, and week-based classification problems respectively. Three significance level are used:

- 99% significance level (p-value is less than 0.01). Symbol '$\Downarrow$ ' corresponds

to 99% significant difference where the mean in users group is less than mean in non-users group and symbol '⇑' corresponds to 99% significant difference where the mean in users group is greater than the mean in non-users group.

- 98% significance level (p-value is less than 0.02). Symbol '↓ ' corresponds to 98% significant difference where the mean in users group is less than mean in non-users group and symbol '↑' corresponds to 98% significant difference where the mean in users group is greater than the mean in non-users group.

- 95% significance level (p-value is less than 0.05). Symbol '⇃ ' corresponds to 95% significant difference where the mean in users group is less than mean in non-users group and symbol '↿' corresponds to 95% significant difference where the mean in users group is greater than the mean in non-users group.

Empty cells in the tables below correspond to insignificant differences.

For example for the decade-based user/non-user separation (see Table 4.10) chocolate does not have a significant difference between users and non-users for any of the factors. Alcohol users and non-users only have a 99% significant difference in the C , Imp, and SS scores, and 95% significant difference in the A score. According to Table 4.6 all these deviations are small.

LSD and magic mushrooms for the decade-based user/non-user separation (see Table 4.10) have 99% significant difference between users and non-users in the O, A, C, Imp, and SS scores and 95% significant difference in the N score. According to Table 4.6 both for LSD and magic mushrooms the deviation in the N score is small and all the deviations in the O, A, C, Imp, and SS scores are not small.

Benzodiazepines and methadone in Table 4.10 have 99% significant differences between users and non-users in all seven scores. For methadone, all these differences are not small (Table 4.6) and for benzodiazepines the difference in the E score is small despite of it 99% significance.

The significance of the differences of the means for groups of users and non-users

Table 4.10. Significant differences of means for groups of users and non-users for the decade-based user/non-user separation.

Drug	N	E	O	A	C	Imp	SS
Chocolate							
Alcohol				⇃	⇓	⇑	⇑
Amyl nitrite			⇑	⇓	⇓	⇑	⇑
Caffeine		↿	⇑		↓	↑	⇑
LSD, Mushrooms	↿		⇑	⇓	⇓	⇑	⇑
Amphetamine, Cocaine, Crack, Ecstasy, Ketamine, Legal highs, Nicotine, VSA	⇑		⇑	⇓	⇓	⇑	⇑
Cannabis, Heroin	⇑	⇃	⇑	⇓	⇓	⇑	⇑
Benzodiazepines, Methadone	⇑	⇓	⇑	⇓	⇓	⇑	⇑
Benzodiazepines pleiad, Heroine pleiad	⇑		⇑	⇓	⇓	⇑	⇑
Ecstasy.pleiad	⇑	⇃	⇑	⇓	⇓	⇑	⇑
Illicit drugs	⇑	⇓	⇑	⇓	⇓	⇑	⇑

for the year, month, and week-based user definition is presented in Tables 4.11–4.13). We hope that the previous descriptions of where the significant differences lie are enough for the reader to interpret this table. It is useful to consider significance of differences together with their value (tables 4.7– 4.9). Additional information about these differences could be extracted from the detailed tables C.1–C.4 in Appendix.

Mean values of five factor scores for groups of drug users and non-users for the decade-based user/non-user separation are depicted in Fig 4.6. A single drug was chosen to be plotted for each group, due to the fact that the shapes of the profile for all drugs in one group are very similar. Fig 4.6 represents T-score graphs of the mean of personality factor scores for the groups of users and non-users with respect to the population norm mean (the left column) and with respect to the sample mean (the right column) for alcohol, LSD, cannabis, and heroin. Graphs of the same type are presented for the year-based classification problem for ketamine in Fig 4.7, for amyl nitrite for the month-based classification problem in Fig 4.8 and for crack for the week-based classification problem in Fig 4.9. Mean values of all seven factor scores for groups of drug users and non-users for all definitions of users are presented in Tables C.2, C.3 and C.4.

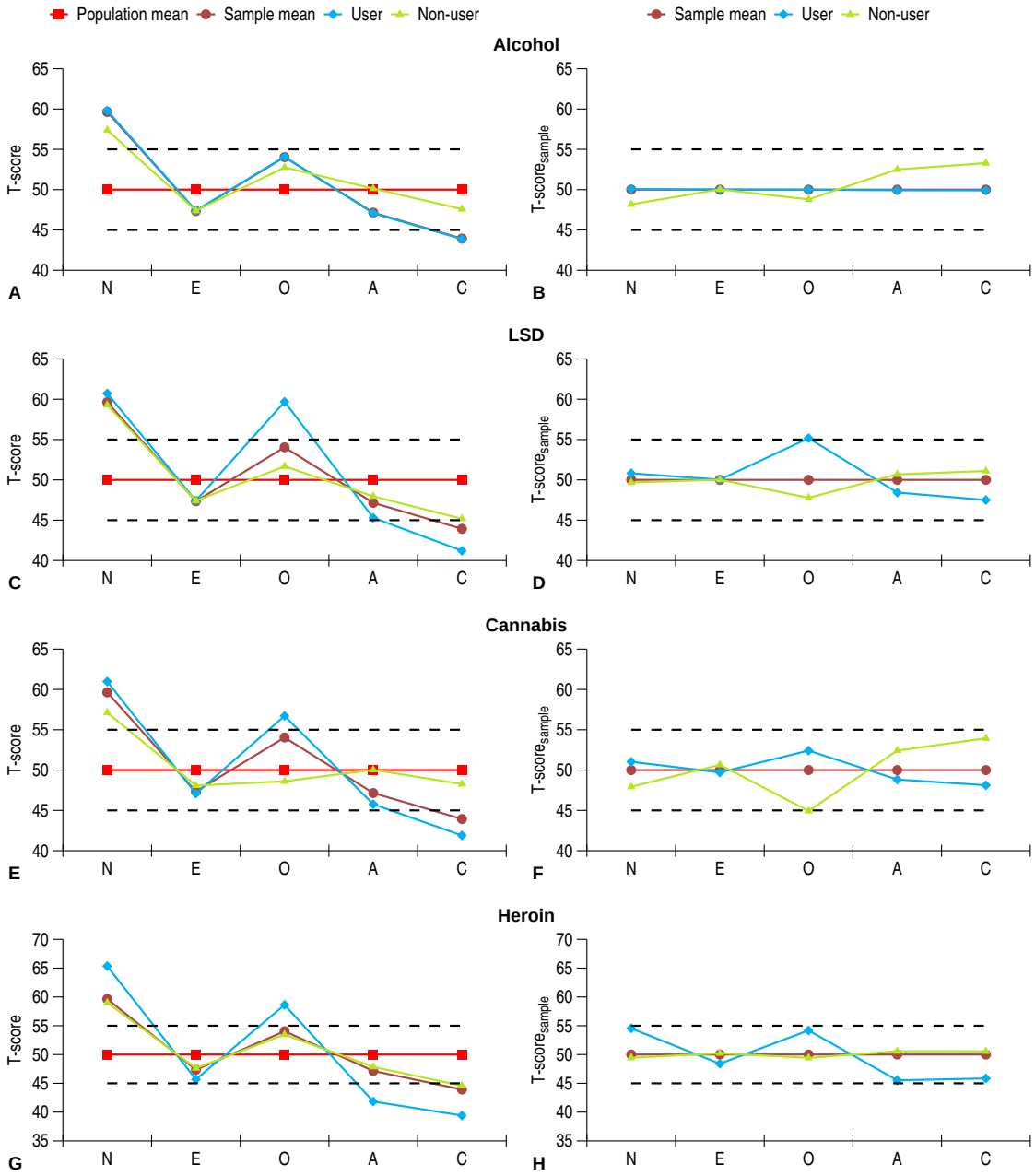

Figure 4.6. Average personality profiles for the decade-based user/non-user separation. T-scores with respect to the population norm mean (left column) and T-score$_{sample}$ with respect to the sample means (right column) for: A & B: Alcohol, C & D: LSD, E & F: Cannabis, and G & H: Heroin

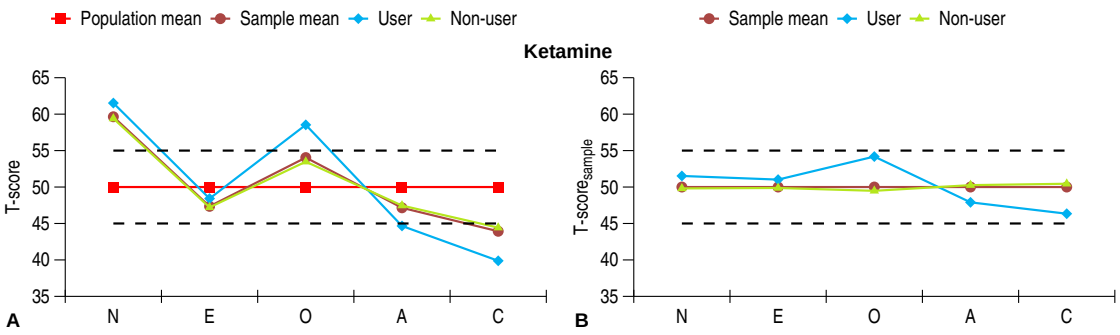

Figure 4.7. Average personality profiles for Ketamine for the year-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means

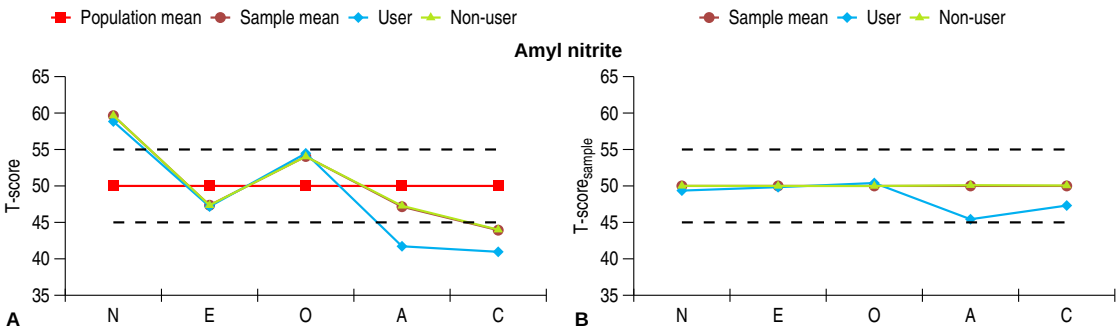

Figure 4.8. Average personality profiles for Amyl nitrite for the month-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means

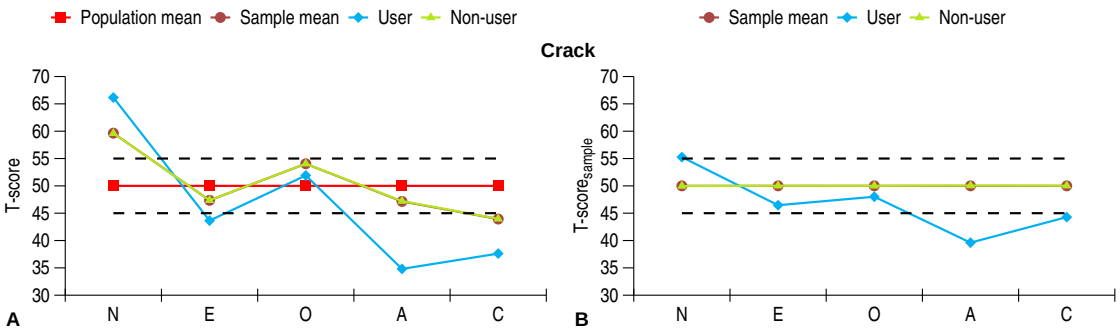

Figure 4.9. Average personality profiles for Crack for the week-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means

Table 4.11. Significant differences of means for groups of users and non-users for the year-based user/non-user separation.

Drug	N	E	O	A	C	Imp	SS
Chocolate							
Alcohol							⇑
Amyl nitrite			↑	⇓	⇓	⇑	⇑
LSD			⇑	↓	⇓	⇑	⇑
Mushrooms			⇑	⇓	⇓	⇑	⇑
Caffeine		↑	↑		↓	⇑	⇑
Ecstasy		⇑	⇑	⇓	⇓	⇑	⇑
Ketamine	↑		⇑	⇓	⇓	⇑	⇑
VSA	⇑		⇑	↓	⇓	⇑	⇑
Amphetamine, Cannabis, Crack, Legal highs, and Nicotine	⇑		⇑	⇓	⇓	⇑	⇑
Cocaine	⇑	↑	⇑	⇓	⇓	⇑	⇑
Heroin	⇑	↓	⇑	⇓	⇓	⇑	⇑
Benzodiazepines, Methadone	⇑	⇓	⇑	⇓	⇓	⇑	⇑
Heroine pleiad	⇑		⇑	⇓	⇓	⇑	⇑
Benzodiazepines pleiad, Ecstasy pleiad	⇑	↓	⇑	⇓	⇓	⇑	⇑
Illicit drugs	⇑	⇓	⇑	⇓	⇓	⇑	⇑

Table 4.12. Significant differences of means for groups of users and non-users for the month-based user/non-user separation.

Drug	N	E	O	A	C	Imp	SS
Chocolate							
Amyl nitrite				↓	↓	↑	⇑
LSD, Mushrooms			⇑		⇓	⇑	⇑
VSA			⇑	↓		⇑	⇑
Ketamine			⇑	↓	⇓	⇑	⇑
Caffeine		⇑				⇑	⇑
Alcohol		⇑	↑				⇑
Ecstasy		⇑	⇑	⇓	⇓	⇑	⇑
Crack	⇑			↓	↓	↑	⇑
Amphetamine, Cannabis, Cocaine, Legal highs, and Nicotine	⇑		⇑	⇓	⇓	⇑	⇑
Heroin	⇑	↓		⇓	⇓	⇑	⇑
Benzodiazepines, Methadone	⇑	⇓	⇑	⇓	⇓	⇑	⇑
Ecstasy pleiad, Heroine pleiad	⇑		⇑	⇓	⇓	⇑	⇑
Benzodiazepines pleiad	⇑	⇓	⇑	⇓	⇓	⇑	⇑
Illicit drugs	⇑	⇓	⇑	⇓	⇓	⇑	⇑

Table 4.13. Significant differences of means for groups of users and non-users for the week-based user/non-user separation.

Drug	N	E	O	A	C	Imp	SS
Chocolate							↓
Crack				⇃		↿	⇑
Amyl nitrite				⇃	⇃	↿	⇑
Ketamine			↑	↓	↓	↑	⇑
VSA			⇑			⇑	⇑
LSD			⇑		⇓	↿	⇑
Cannabis			⇑	⇓	⇓	⇑	⇑
Mushrooms		↿	⇑			⇑	⇑
Caffeine		↑				↿	
Ecstasy		↑	⇑		⇓	⇑	⇑
Alcohol		⇑					⇑
Cocaine	↑			⇓	⇓	⇑	⇑
Amphetamine, Nicotine	⇑		⇑	⇓	⇓	⇑	⇑
Heroin	⇑	⇃		⇓	⇓	⇑	⇑
Methadone	⇑	⇓	↿	⇓	⇓	⇑	⇑
Benzodiazepines, Legal highs	⇑	⇓	⇑	⇓	⇓	⇑	⇑
Ecstasy pleiad	⇑		⇑	⇓	⇓	⇑	⇑
Benzodiazepines pleiad, Heroine pleiad	⇑	⇓	⇑	⇓	⇓	⇑	⇑
Illicit drugs	⇑	⇓	⇑	⇓	⇓	⇑	⇑

4.6 Correlation between usage of different drugs

Usage of each drug is a binary variable (users or non-users) for all versions of user definition. Tables C.5 and C.6 contain PCCs, which are computed for each pair of the 153(=18 times 17 divided by 2) potential drug usages for the decade- and year-based user/non-user separations respectively (see Appendix C.2). The majority of the PCCs are significant, since the sample size is 1885.

The correlation coefficient is an indicator of measuring the dependence between attributes. The Pearson's Correlation Coefficient (PCC) can be defined as the covariance of two random variables divided by the product of the individual standard deviations. Let as consider two variables X and Y, for a series of n measurements of X and Y written x_i and y_i, where $i = 1, 2, \ldots, n$. The Pearson's r is defined as:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean, and analogously for $\bar{y}$. r is a value between $-1 \leq r \leq 1$, where -1 is total negative correlation, 1 is total positive correlation, and 0 is no correlation.

PCC is not appropriate for general categorical attributes but for the Boolean random variables (with 0,1 values) it gives a reasonable measure of dependence because for them $cov(X, Y) = P\big(X = 1 \& Y = 1\big) - P\big(X = 1\big)P\big(Y = 1\big)$.

The correlation in 124 pairs of drug usages from a totality of 153 pairs for the decade-based classification problem have p-values less than 0.01 (p-value is the probability to observe by chance the same or greater correlation coefficient for uncorrelated variables). It is necessary to employ a *multi-testing* approach when testing 153 pairs of drug usages in order to estimate the significance of the correlation [138]. We apply the most conservative technique, the Bonferroni correction, and used the Benjamini-Hochberg (BH) step-up procedure [138] to control the False Discovery Rate (FDR) in order to estimate the genuine significance of these correlations. The FDR is the expected proportion of false positives among all discoveries (rejected null hypotheses). Let as consider the problem of m simultaneously tested null hypothesis of which m_0 are true. For each hypothesis $H_i\big(i = 1, \ldots, m\big)$ a test statistic is calculated along with corresponding p-value $P_1, P_2, \ldots, P_m$. The FDR defined as

$$FDR = E\Big[\frac{F}{F+T}\Big] = E\Big[\frac{F}{R}\Big],$$

where F is the number of false positives (false discoveries), T is the number of true positives (true discoveries), R is the number of rejected null hypotheses (discoveries).

Consider $H_1, H_2, \ldots, H_m$ be a family of null hypotheses tested based on the corresponding p-value $P_1, P_2, \ldots, P_m$. Let as consider the increasing order p-value

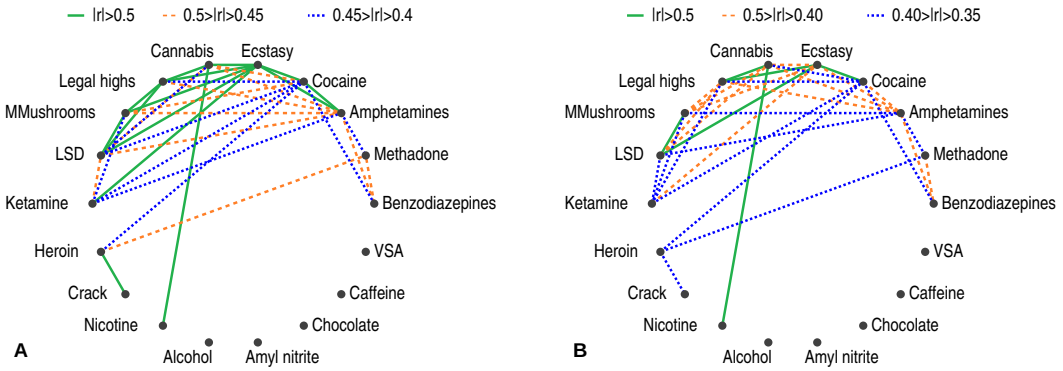

Figure 4.10. Strong drug usage correlations: A: for the decade-based classification problem and B: for the year-based classification problem

denoted by $P_{(1)} < P_{(2)} < \ldots < P_{(m)}$ and $H_{(i)}$ is the null hypothesis with respect to $P_{(i)}$. Then Bonferroni procedure works as follows: For a given α, find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$. Then reject all null hypothesis $H_{(i)}, i = 1, 2, \ldots, k$. The Benjamini-Hochberg procedure (BH step-up procedure) controls the FDR (at level α). The procedure works as follows: For a given α, find the largest k such that $P_{(k)} \leq \frac{k}{m+1-k}\alpha$. Then reject all null hypothesis $H_{(i)}, i = 1, 2, \ldots, k$.

There are 115 significant correlation coefficients with Bonferroni corrected p-value 0.001. The BH step-up procedure with threshold of FDR equal to 0.01 defines 127 significant correlation coefficients.

However, a significant correlation does not necessarily imply a strong association or causality. For example, the correlation coefficient for alcohol usage and amyl nitrate usage is significant (i.e. the p-value is equal to 0.0013) but the value of this coefficient is equal to 0.074, and thus cannot be considered as an important association. We consider correlations with absolute values of PCC $|r| \geq 0.4$. Fig 4.10 sets out all significant identified correlations greater than 0.4. In this study for the decade-based classification problem we consider the correlation as weak if $|r| < 0.4$, medium if $0.45 > |r| \geq 0.4$, strong if $0.5 > |r| \geq 0.45$, and very strong if $|r| \geq 0.5$.

The correlation coefficient is high for each pair from the group: amphetamines, cannabis, cocaine, ecstasy, ketamine, legal highs, LSD, and magic mushrooms,

excluding correlations between cannabis and ketamine usage (r=0.302) and between legal highs and ketamine usage (r=0.393) (Fig 4.10A). Crack, benzodiazepines, heroin, methadone, and nicotine usages are correlated with one, two, or three other drugs usage (see Fig 4.10A). Amyl nitrite, chocolate, caffeine and VSA usage are uncorrelated or weakly correlated with usage of any other drug.

The structure of correlations of the year-based user/non-user separation is approximately the same as for the decade-based classification problem (see Fig 4.10). We consider correlations with absolute values of PCC $|r| \geq 0.35$ for the year-based classification. Fig 4.10B sets out all identified significant correlations with $|r| > 0.35$. The correlation can be interpreted as weak if $|r| < 0.35$; medium if $0.40 > |r| \geq 0.35$; strong if $0.5 > |r| \geq 0.40$; and very strong if $|r| \geq 0.5$. On base of this similarity of correlation structures we define pleiades for three central drugs: heroin, ecstasy, and benzodiazepines (as described in the Section 'Pleiades of drugs').

Relative Information Gain (RIG) is widely used in data mining to measure the dependence between categorical attributes [139]. RIG of the drug X usage with respect to the drug Y usage is defined as:

$$RIG(X|Y) = \frac{(IG(X|Y))}{Entropy(X)} = \frac{(Entropy(X) - Entropy(X|Y))}{Entropy(X)},$$

where $Entropy(X)$ is the entropy of drug X usage:

$$Entropy(X) = -\mu ln\mu - (1-\mu)ln(1-\mu),$$

where μ is the fraction of drug X users among all participants, and $Entropy(X|Y)$ is the relative entropy:

$$Entropy(X|Y) = \nu Entropy(X|y = User) + (1-\nu)Entropy(X|y = Non-user),$$

where ν is the fraction of drug Y users among all participants, $Entropy(X|y =$

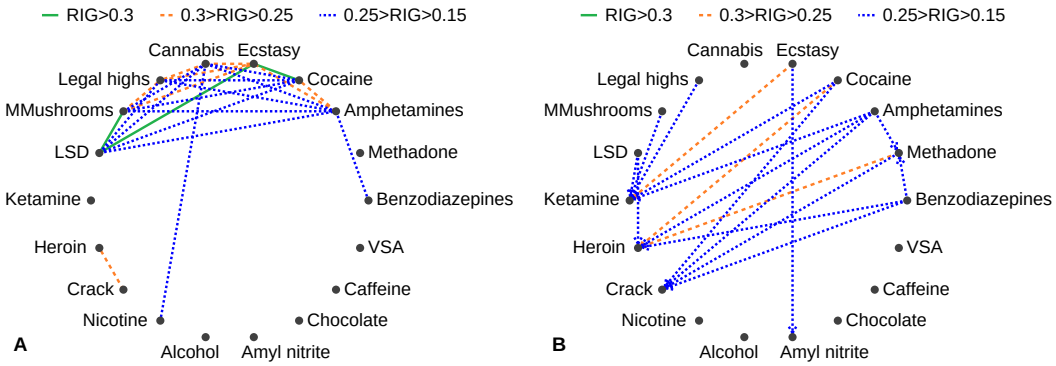

Figure 4.11. Pairs of decade-based drug usages with high RIG: A: approximately symmetric RIG and B: significantly asymmetric RIG. In figure B the arrow from cocaine usage to heroin usage, for example, means that knowledge of cocaine usage can decrease uncertainty in knowledge about heroin usage.

$User)$ and $Entropy(X|y = Non - user)$ are the specific conditional entropies:

$$Entropy(X|y = User) = -\mu_{(y=User)} ln \mu_{(y=User)} - (1 - \mu_{(y=User)}) ln(1 - \mu_{(y=User)}),$$

$$Entropy(X|y = Non - user) = -\mu_{(y=Non-user)} ln \mu_{(y=Non-user)} -$$

$$(1 - \mu_{(y=Non-user)}) ln(1 - \mu_{(y=Non-user)}),$$

where $\mu_{(y=User)}$ is the fraction drug X user among all drug Y users and $\mu_{(y=Non-user)}$ is the fraction of drug X users among all drug Y Non-users.

The greater the value of RIG is, the stronger the indicated correlation is. RIG is zero for independent attributes, is not symmetric and is a measure of mutual information. For example, the value of RIG for drug 1 usage from usage of drug 2 is equal to a fraction of uncertainty (entropy) in drug 1 usage, which can be removed if the value of drug 2 usage is known. The significance of RIG for binary random variables is the same as for PCC. The majority of RIGs are significant, but have small values. Fig 4.11 presents all pairs with RIG >0.15.

Fig 4.11A shows 'approximately symmetric' RIGs. Here, we call $\text{RIG}(X|Y)$ approximately symmetric if

$$\frac{|\text{RIG}(X|Y) - \text{RIG}(Y|X)|}{\min(\text{RIG}(X|Y), \text{RIG}(Y|X))} < 0.2.$$

RIG is approximately symmetric for each pair from the following group: am-

Table 4.14. The results of feature ranking. Data include country of residence and ethnicity quantified by CatPCA. FVE is the fraction of explained variance. CFVE is the cumulative FVE. The least informative features are located towards the bottom of the table.

Principal variable ranking			Double Kaiser's ranking
Attribute	FVE	CFVE	
SS	0.192	0.192	E
N	0.153	0.345	C
A	0.106	0.451	SS
Edu	0.104	0.555	N
O	0.092	0.647	Imp
C	0.088	0.735	O
E	0.076	0.811	A
Age	0.073	0.884	Age
Imp	0.055	0.939	Edu
Country	0.037	0.976	Country
Gndr	0.021	0.997	Gndr
Ethnicity	0.003	1.000	Ethnicity

phetamines, cannabis, cocaine, ecstasy, legal highs, LSD and magic mushrooms. This group is the same that in Fig 4.10 (except ketamine). Fig 4.11B shows asymmetric RIGs. Asymmetric RIGs illustrate a markedly different pattern to that of Fig 4.10.

4.7 Input feature ranking

It should be stressed that the five factor model (FFM), impulsivity, and sensation-seeking are all correlated. To identify the most informative features we apply the methods which are described in the Section 'Input feature ranking' in chapter 3. The results of the principal variables calculation are given in Table 4.14 for CatPCA quantification, and in Table 4.15 for the dummy coding of nominal features. Tables 4.14 and 4.15 contain lists of attributes in order from best to worst. The results of the double Kaiser ranking are shown in the same tables.

The results of application of sparse PCA are shown in Tables 4.16 and 4.17. As a result of feature selection we can exclude ethnicity from further consideration. There is a more intriguing effect regarding country of location. Only two coun-

Table 4.15. The results of feature ranking. Data include dummy coded country of residence and ethnicity. FVE is the fraction of explained variance. CFVE is the cumulative FVE. The least informative features are lower located.

Principal variable ranking			Double Kaiser's ranking
Attribute	FVE	CFVE	
SS	0.186	0.186	E
N	0.149	0.335	C
A	0.103	0.438	SS
Edu	0.101	0.539	N
O	0.089	0.627	Imp
C	0.086	0.714	O
E	0.074	0.787	A
Age	0.071	0.858	Age
Imp	0.053	0.911	Edu
UK	0.027	0.938	UK
Gndr	0.020	0.959	USA
USA	0.013	0.972	Gndr
White	0.010	0.982	Other (country)
Other (country)	0.005	0.988	White
Canada	0.004	0.991	Other (ethnicity)
Other (ethnicity)	0.003	0.994	Canada
Black	0.002	0.995	Asian
Australia	0.002	0.997	Mixed-White/Black
Asian	0.001	0.998	Australia
Mixed-WhiteBlack	0.001	0.999	Black
Republic of Ireland	0.000	1.000	Mixed-White/Asian
Mixed-WhiteAsian	0.000	1.000	Republic of Ireland
New Zealand	0.000	1.000	New Zealand
Mixed-BlackAsian	0.000	1.000	Mixed-Black/Asian

tries are informative (in this sample): the UK and the USA. Furthermore, inclusion of country in the personality measures does not add much to the prediction of drug usage. To understand the reasons for these two countries' importance in the prediction of drug consumption we compare the statistics for the subsamples: UK - non-UK and USA - non-USA. We calculated the p-value for coincidence of distribution of personality measurements in each subsample. We obtained the same results for both divisions into subsamples: all input features have significantly different distributions with a 99.9% confidence level for UK and non-UK subsamples and likewise for USA – non-USA subsamples. This means that the

Table 4.16. The result of sparse PCA feature ranking. Data include country of residence and ethnicity quantified by CatPCA.

Step	# of components	Removed attributes
1	5	Gndr and Ethnicity
2	4	No removed attributes. The retained set of attributes: age, Edu, N, E, O, A, C, Imp, SS, and country

Table 4.17. The result of sparse PCA feature ranking. Data include dummy coded country of residence and ethnicity.

Step	# of components	Removed attributes
1	8	Canada, Other (country), Australia, Republic of Ireland, New Zealand, Mixed-White/Asian, White, Other (ethnicity), Mixed-White/Black Asian, Black and Mixed-Black/Asian
2	5	Gndrr, UK and USA
3	4	No removed attributes. The retained set of attributes: age, Edu, N, E, O, A, C, Imp, and SS

UK and non-UK samples are biased, and similarly for the USA and non-USA samples.

Our goal is to predict the risk of drug consumption for an individual. This means that we have to consider individual specific factors. Occupation within a specific country can be thought of as an important risk factor, but we do not have enough data for countries other than the UK and the USA because of the composition of the dataset: 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 (20; 1.1%). A total of 118 (6.3%) came from a diversity of other countries, none of whom individually formed as much as 1% of the sample, or did not declare the country of location. Thus we exclude the 'country' feature from further study. As a result, we continue with the 10 input features: Age, Edu, N, E, O, A, C, Imp, SS, and Gndr.

4.8 Selection of the best classifiers for the decade-based classification problem

The first step for risk evaluation is the construction of classifiers. We have tested the eight methods described in the 'Risk evaluation methods' Section and selected the best one. The results of classifier selection are presented in Table 4.18. This table shows that for all drugs except alcohol, cocaine and magic mushrooms, the sensitivity and specificity are greater than 70%, which is an unexpectedly high accuracy.

Recall that we have 10 input features: Age, Edu, N, E, O, A, C, Imp, SS, and Gndr; each of which is an important predictor for at least five drugs. However, there is no single most effective classifier which uses all input features. The maximal number of attributes used is 6 out of 10 and the minimal number is 2. In Section 'Criterion of the best method selection' the best method is defined as the method which maximises value of the minimum of sensitivity and specificity. If the minimum of sensitivity and specificity is the same for two classifiers then the classifier with the maximal sum of the sensitivity and specificity is selected from these. Table 4.18 shows the different sets of attributes which are used in the best user/non-user classifier for each different drug.

The use of a features in the best classifier can be interpreted as 'ranking by fact'. We note that this ranking by fact is very different from the other rankings presented in Tables 4.14 and 4.16. For example, in Tables 4.14 and 4.16 we see that age is not one of the most informative measures, but it is used in the best classifiers for 14 of the drugs. The second most used input feature is Gndr, which is regarded as non-informative by Sparse PCA (Table 4.16) and as one of the least informative by other methods (Table 4.14). This means that consumption of these 10 drugs is Gndr dependent.

We found some unexpected outcomes: for example, in the dataset the fraction of females who are alcohol users is greater than that fraction of males (Fig 4.12)

Table 4.18. The best results of the drug users classifiers (decade-based definition of users). Symbol 'X' means the used input feature. Results are calculated by LOOCV.

Target feature	Classifier	Age	Edu	N	E	O	A	C	Imp	SS	Gndr	Sn (%)	Sp (%)	Sum (%)
Alcohol	LDA	X	X	X						X	X	75.34	63.24	138.58
Amphetamines	DT	X		X		X		X	X	X		81.30	71.48	152.77
Amyl nitrite	DT			X		X		X		X		73.51	87.86	161.37
Benzodiazepines	DT	X		X	X				X	X	X	70.87	71.51	142.38
Cannabis	DT	X	X			X	X	X	X			79.29	80.00	159.29
Chocolate	KNN	X			X			X			X	72.43	71.43	143.86
Cocaine	DT	X				X	X		X	X		68.27	83.06	151.32
Caffeine	KNN	X	X			X	X		X			70.51	72.97	143.48
Crack	DT				X			X				80.63	78.57	159.20
Ecstasy	DT	X								X	X	76.17	77.16	153.33
Heroin	DT	X							X		X	82.55	72.98	155.53
Ketamine	DT	X			X		X		X	X		72.29	80.98	153.26
Legal highs	DT	X				X	X	X		X	X	79.53	82.37	161.90
LSD	DT	X		X	X	X			X		X	85.46	77.56	163.02
Methadone	DT	X	X		X	X					X	79.14	72.48	151.62
MMushrooms	DT				X						X	65.56	94.79	160.36
Nicotine	DT			X	X			X			X	71.28	79.07	150.35
VSA	DT	X	X		X		X	X		X		83.48	77.64	161.12

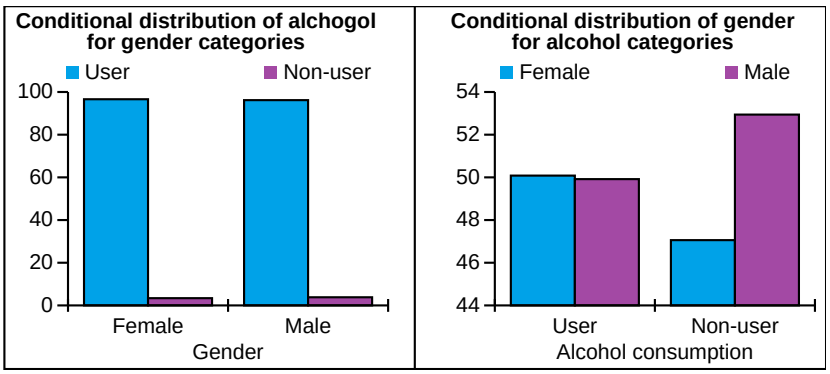

Figure 4.12. Conditional distribution for gender and alcohol.

but a greater proportion of males consume caffeine drinks (for example, coffee) (Fig 4.13). The fraction of males who do not eat chocolate is greater than for females (Fig 4.14). The conditional distributions for nicotine show the fraction of males who smoke is higher (Fig 4.15).

The next most informative input features are E and SS which are used in the best classifiers for nine drugs. Features O, C, and Imp are used in the best classifiers for eight drugs. Features N and A are used in the best classifiers for six drugs. Thus, personality factors are associated with drug use and each one impacts on

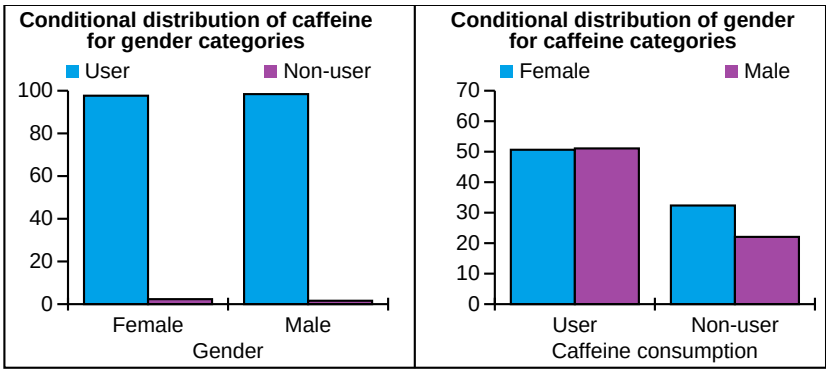

Figure 4.13. Conditional distribution for gender and caffeine.

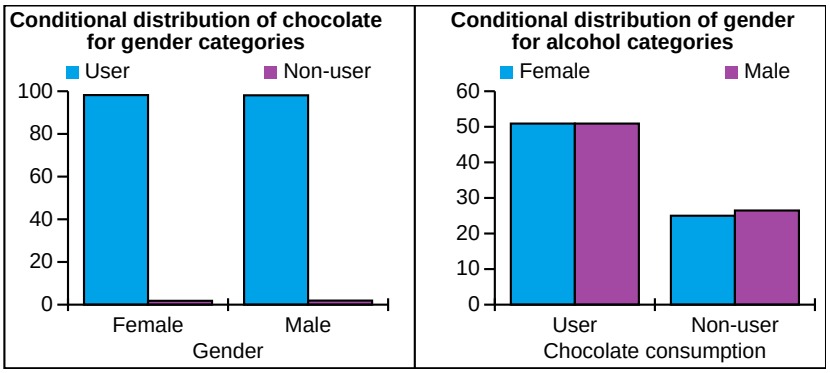

Figure 4.14. Conditional distribution for gender and chocolate.

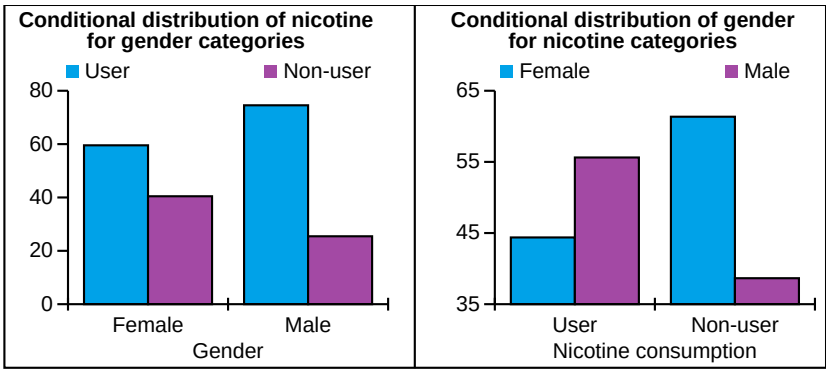

Figure 4.15. Conditional distribution for gender and nicotine.

specific drugs. Finally, Edu is used in the best classifiers for five drugs (see Table 4.18).

To predict the usage of most drugs DT is the best classifier (see Table 4.18). LDA is the best classifier for alcohol use with five input features, and has sensitivity 75.34% and specificity 63.24%. *k*NN is the best classifier for chocolate and caffeine users. These *k*NN classifiers use four features for chocolate and five features for caffeine.

The drugs can be separated into disjoint groups by the number of attributes used for the best classifiers:

- The group of classifiers with two input features contains classifiers for two drugs: crack and magic mushrooms. Both classifiers in this group use the E score.
- The group of classifiers with three input features includes classifiers for two drugs: ecstasy and heroin. Both classifiers in this group use Age and Gndr and do not use any NEO-FFI factors.
- The group of classifiers with four input features includes classifiers for three drugs: amyl nitrite, chocolate, and nicotine. All classifiers in this group use the C score.
- The group of classifiers with five input features includes classifiers for five drugs: alcohol, cocaine, caffeine, ketamine, and methadone. All classifiers in this group use age.
- The group of classifiers with six input features includes classifiers for six drug users: amphetamines, benzodiazepines, cannabis, legal highs, LSD, and VSA. All classifiers in this group use Age.

It is important to stress that the *attributes which are not used in the best classifiers are not non-informative*. For example, for ecstasy consumption the best classifier is based on age, SS, and Gndr and has sensitivity 76.17% and specificity 77.16%. There exists a DT for usage of the same drug based on Age, Edu, O, C, and SS, with sensitivity 77.23% and specificity 75.22%; a DT based on Age, Edu, E, O, and A, with sensitivity 73.24% and specificity 78.22%; a LR classifier based on Age, Edu, O, C, Imp, SS, and Gndr, with sensitivity 74.83% and specificity 74.52%; a *k*NN classifier based on Age, Edu, N, E, O, C, Imp, SS, and Gndr, with sensitivity 75.63% and specificity 75.75%. This means that for the risk evaluation of ecstasy usage all input attributes are informative but the required information can be extracted from a smaller subset of the attributes.

Table 4.19. The best results of the drug users classifiers in the space of the first four principal components. Symbol 'X' means used input feature. Results are calculated by LOOCV.

Target feature	Classifier	PC 1	PC 2	PC 3	PC 4	Sn (%)	Sp (%)	Sum (%)
Alcohol	GM	X	X			54.71	70.59	125.29
Amphetamines	DT	X				74.37	77.78	152.15
Amyl nitrite	DT		X		X	61.35	79.47	140.82
Benzodiazepines	DT			X		64.63	92.20	156.83
Cannabis	DT	X	X			78.02	77.26	155.28
Chocolate	LDA		X	X	X	57.35	62.86	120.21
Cocaine	DT				X	72.20	85.23	157.42
Caffeine	LDA	X	X	X		62.55	78.38	140.93
Crack	DT	X	X			78.01	77.10	155.11
Ecstasy	DT	X	X		X	73.10	73.46	146.56
Heroin	DT	X				76.89	74.72	151.60
Ketamine	DT			X		73.43	93.55	166.98
Legal highs	PDFE	X	X	X	X	75.98	76.14	152.12
LSD	DT		X			99.46	61.30	160.76
Methadone	DT			X	X	79.38	84.47	163.85
MMushrooms	LDA	X	X	X	X	76.37	69.10	145.47
Nicotine	DT	X		X	X	79.67	74.56	154.23
VSA	DT		X	X		83.48	84.23	167.71

Sets of informative input features for each drug usage in the space of the first four PCs are presented in Table 4.19.

The results presented in Tables 4.18 and 4.19 were calculated by LOOCV. It should be stressed that different methods of testing give rise to different values for sensitivity and specificity. Common methods include calculation of test set errors (the holdout method), k-fold cross-validation, testing on the entire sample (if it is sufficiently large, the so-called 'naïve' method), random sampling, and many others. For example, a DT formed for the entire sample can have a sensitivity and specificity different from LOOCV [117]. For illustration, consider the DT for ecstasy, depicted in the Fig 4.16. It has sensitivity 78.56% and specificity 71.16%, calculated using the whole sample. The results of LOOCV for a tree with the same choices are given in Table 4.18: sensitivity 76.17% and specificity 77.16%.

The role of SS is very important for most of the party drugs. In particular, the risk of ecstasy consumption can be evaluated with high accuracy on the basis of age, Gndr and SS (see Table 4.18, Fig 4.16, and 4.34), and does not need the personality traits from the FFM.

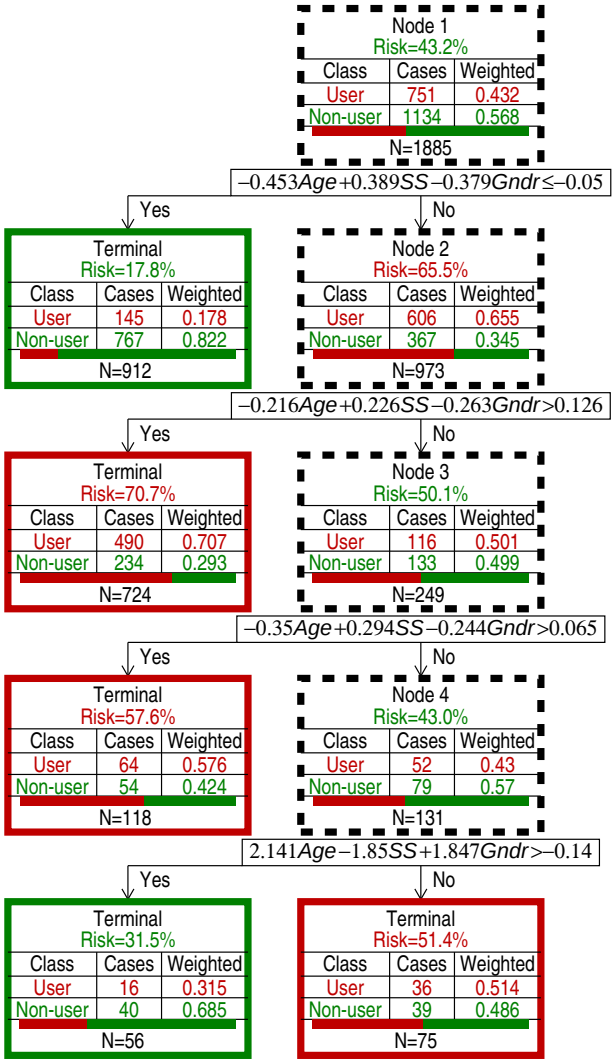

Figure 4.16. Decision tree for ecstasy. Input features are: Age, SS, and Gndr. Non-terminal nodes are depicted with dashed border. Values of Age, SS, and Gndr are calculated by quantification procedures described in the 'Input feature transformation' Section. The weight of each case of users class is 1.15 and of non-users class is 1. Column 'Weighted' records normalized weights: the weight of each class is divided by sum of weights.

4.9 The best class binarization

We have seven categories for drug users, but to classify them into seven classes we can only use the following classifiers: KNN, DT, RF and PDFE. All other classifier implementation is appropriate for binary classification only. For this purpose, we should change the seven order classes into binary class. In addition, binary classification is a better understood task and is simpler because only two classes

Table 4.20. Possible class binarization

Name of binarization	The first class	The second class
Strongly never	Never used	Used over a decade ago, Used in last decade, Used in last year, Used in last month, Used in last week, Used in last day
User/non-user	Never used, Used over a decade ago	Used in last decade, Used in last year, Used in last month, Used in last week, Used in last day
Used long ago	Never used, Used over a decade ago, Used in last decade	Used in last year, Used in last month, Used in last week, Used in last day
About year	Never used, Used over a decade ago, Used in last decade, Used in last year	Used in last month, Used in last week, Used in last day
About month	Never used, Used over a decade ago, Used in last decade, Used in last year, Used in last month	Used in last week, Used in last day
Frequently	Never used, Used over a decade ago, Used in last decade, Used in last year, Used in last month, Used in last week	Used in last day

are involved.

We have seven ordered classes. There are six possible division of these set of classes into two classes without changing the order of classes. The best possible class, which is binarization, is represented in Table 4.20. We were testing all methods to define the best class binarization for original input features and space of the first four principal components for each drug.

We can see that for different target attribute we have six different class binarization: (Strongly never, User/non-user, Used long ago, About year, About month, Frequently). Recall that we have two space: space of original input feature and space of the first four PCs. We applied eight data mining methods to predict class users for all drugs and for each six class binarization in both spaces.

The aim is to select the best classifiers to each class binarization for all drugs in each space. For this purpose we chose the best criteria which we employ the sum of sensitivity and specificity. After that we select the best classifier among all classifiers and we obtain the best class binarization in each space for all drugs. For example, ecstasy: in the space of original attribute the best classifier for class

binarization strongly never is KNN, user/non-user is GM, used long ago is DT, about year is DT, about month is LDA and Frequently is LR. Table 4.21 represents the best classifiers to each class binarization in the space of original input features for ecstasy usage.

Table 4.21. The best classifiers to each class binarization for ecstasy usage in the space of original input features.

Target attribute	Class binarization	classifier	Sp (%)	Sn (%)	Sum %
Ecstasy	Strongly never	KNN	66.60	77.20	143.80
Ecstasy	User/non-user	GM	70.19	79.76	149.95
Ecstasy	Used long ago	DT	76.02	76.02	152.04
Ecstasy	About year	DT	77.45	74.17	151.61
Ecstasy	About month	LDA	64.02	82.14	146.16
Ecstasy	Frequently	LR	71.78	57.14	128.92

By using the criterion we found that DT is the best classifier among all classifiers for ecstasy and we get the best class binarization used long ago for this space. The specificity for this classifier is 76.02%, sensitivity is 76.02% and sum of them is 152.04%. As a result, the best class binarization for ecstasy user is used long ago.

In the space of original input features for every day ketamine user, every day mmushrooms user, and every day VSA user we see that there is no classifier which can separate with appropriate accuracy (sensitivity and specificity both are greater than 50%). It is also true for amyl nitrite user, and chocolate users who never use chocolate or have used it more than a year ago. Moreover, NB classifier allows the exact isolation of every day crack users for both space.

Let us consider the VSA users in the space of the two first PCs. Distributions of all points of each class separately are depicted in Figure 4.17. We can see that all classes from Used in last decade till Used in last day are shifted to left side of figures. It allows to create classifier to recognise this cases.

For different drugs we have the different binary classes which are separated with the best accuracy. For the most of drug the classifier formed in the space of the original attributes and in the space of the first four principal components have the same best class binarizations or the nearest best binarizations. For example, in

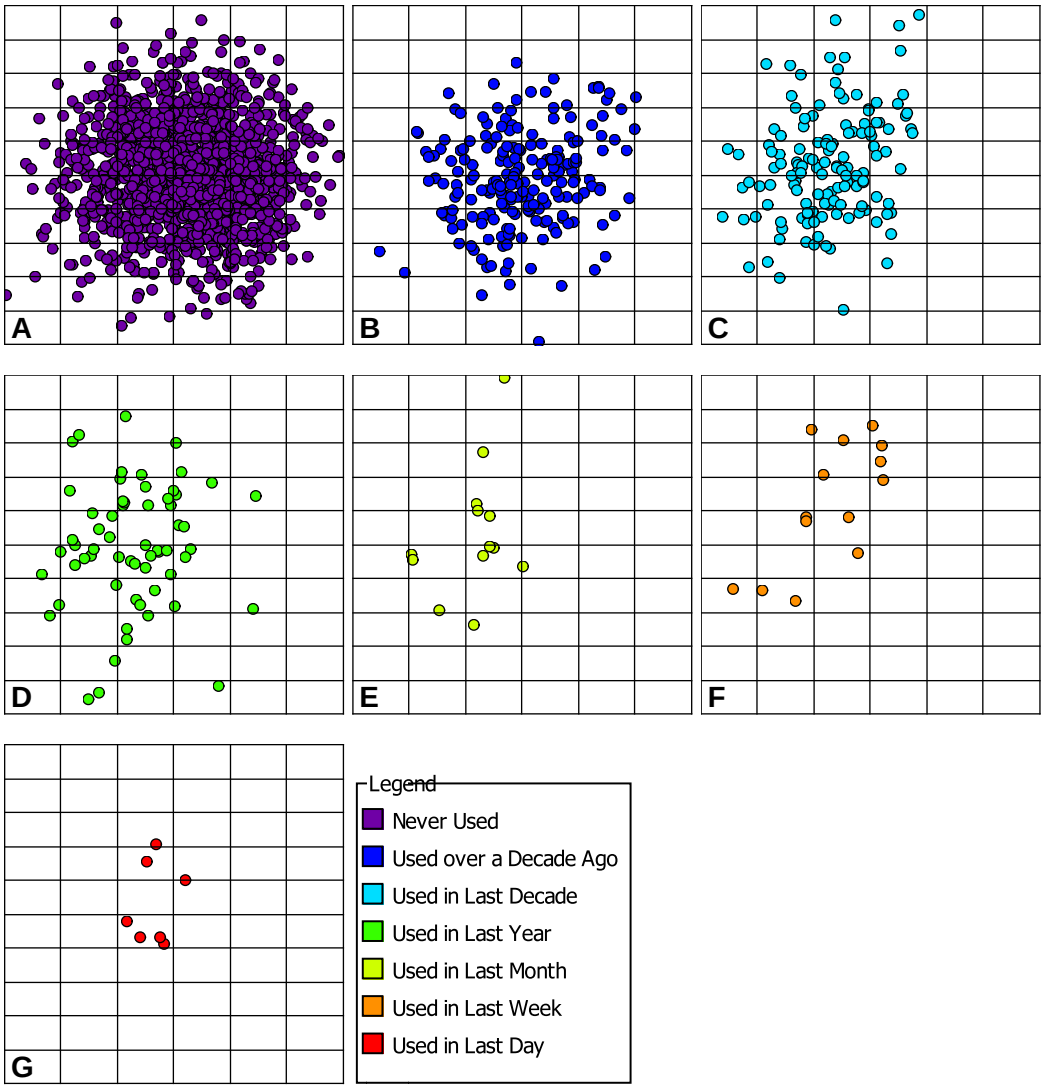

Figure 4.17. The distribution of VSA users in space of the two first principal components: A) Never used, B) Used over a decade ago, C) Used in last decade, D) Used in last year, E) Used in last month, F) Used in last week, G) Used in last day, H) legend.

space of original attribute the best class binarization for alcohol use is about year. The target attribute appears with DT. The true negative rate and true positive rate is 66.17% and 66.67% respectively and the sum of them is 132.83%. In space of PCs the best class binarization for alcohol use is used long ago with the same classifier DT. Sensitivity is 87.82% and specificity is 59.56% and the sum of them is 147.38%.

For cannabis use the best class binarization in space of original attribute is used long ago which appears with RF. Sensitivity for this classifier is 80.95% and speci-

ficity is 80.02% and sum of sensitivity and specificity is 160.98%. Also we can see that the best class binarization in space of PC for cannabis use is used long ago which occur with DT. Specificity for this model is 79.12% and sensitivity is 79.38% and the sum of them is 158.50% . Furthermore, from this example we found that for alcohol use the classifier formed in both space has the nearest best class binarizations. For cannabis use the classifier formed in both space has the same best class binarizations. Table 4.22 and Table 4.23 show the best class binarization in space of original attributes and in space of PCs with the best classifiers for alcohol and cannabis.

Table 4.22. The best class binarization in space of original attributes with the best classifiers for alcohol and cannabis. Sn is sensitivity. Sp is specificity.

Target attribute	Space of original attributes				Classifier
	Class binarization	Sp (%)	Sn (%)	Sum (%)	
Alcohol	About year	66.17	66.67	132.83	DT
Cannabis	Used long ago	80.03	80.95	160.98	RF

Table 4.23. The best class binarization in space of principal components with the best classifiers for alcohol and cannabis. Sn is sensitivity. Sp is specificity.

Target attribute	Space of original attributes				Classifier
	Class binarization	Sp (%)	Sn (%)	Sum (%)	
Alcohol	Used long ago	59.56	87.82	147.38	DT
Cannabis	Used long ago	79.12	79.38	158.50	DT

We compared all the results between the two spaces for each drug. We select the best methods between the two spaces for each drug by the criteria to get the best class binarization. The best class binarization and the best Space with the best classifiers for each drug represented in Table 4.24.

We use classifier of drug users to predict the risk of drug consumer. The relation between personality measures and drug use depends on the types of drugs and frequency of using drugs. Different frequency and different types of drugs have different classifiers to predict the class of the users. The classifiers for users of the most of legal drugs are allowed with the first three class binarization (strongly never, user/non-user and used long ago). Some types of illicit drugs have direct effects on personality traits. A good illustration for this is consumption of

Table 4.24. The best class binarization and the best space with the best classifier. Sn is sensitivity. Sp is specificity. Clas. means the best classifier.

Target attribute	Binarization	Sn (%)	Sp (%)	Sum(%)	The best space	Clas.
Alcohol	Used long ago	59.56	87.82	147.38	Space of PC	DT
Amphetamine	Used long ago	63.22	85.09	148.31	Original	KNN
Amyl nitrite	Used long ago	65.30	77.44	142.74	Space of PC	DT
Benzodiazepines	Frequently	86.82	100	186.82	Original	DT
Cannabis	Used long ago	80.03	80.95	160.98	Original	RF
Chocolate	About month	70.63	62.01	132.65	Space of PC	DT
Cocaine	About year	76.42	74.84	151.26	Original	DT
Caffeine	User/non-user	64.87	66.45	131.32	Original	LDA
Crack	Frequently	100.00	100.00	200.00	Both Spaces	NB
Ecstasy	Used long ago	76.02	76.02	152.04	Original	DT
Heroin	Frequently	77.72	84.62	162.34	Space of PC	LDA
Ketamine	About month	80.41	64.87	145.28	Original	GM
Legal highs	User/non-user	86.11	75.90	162.01	Original	RF
LSD	Frequently	71.53	92.31	163.84	Original	LDA
Methadone	User/non-user	78.47	71.22	149.70	Space of PC	DT
MMushrooms	Used long ago	66.02	90.78	156.81	Original	GM
Nicotine	Strongly never	77.80	75.91	153.71	Space of PC	DT
VSA	Used long ago	89.61	100.00	189.61	Space of PC	DT

benzodiazepines, heroin, LSD and crack because these drugs are appearing with frequently class binarization. Therefore, illegal drug users are predicted to be more in risk factor than legal drug users on personality traits.

4.10 Correlation pleiades of drugs

Consider correlations between drug usage for the year- and decade-based definitions (Fig 4.10). It can be seen from Fig 4.10 that the structure of these correlations for the year- and decade-based definitions of drug users is approximately the same. We found three groups of strongly correlated drugs, each containing several drugs which are pairwise strongly correlated. This means that drug consumption has a 'modular structure', and we identify three modules:

- Crack, cocaine, methadone, and heroin;
- Amphetamines, cannabis, cocaine, ketamine, LSD, magic mushrooms, legal highs, and ecstasy;
- Methadone, amphetamines, cocaine and benzodiazepines.

This modular structure has a clear representation in the correlation graph, Fig 4.10.

The idea of merging correlated attributes into 'modules' is popular in biology. These modules are called *correlation pleiades* [43–45]. This concept was introduced in biostatistics in 1931 [43]. Correlation pleiades were used in evolutionary physiology for the identification of the modular structures in a variety of contexts [43–46]. Berg [45] presented correlation data from three unspecialized and three specialized pollination species, and proposed that correlation pleiades are clusters of correlated traits. This means that, in the standard approach to clustering, the pleiades do not intersect. The classical clustering methods are referred to as 'hard' or 'crisp' clustering, which means that each data object is assigned to only one cluster. This restriction is relaxed for fuzzy [48] and probabilistic clustering [47]. Such approaches are useful when the boundaries between clusters are not well separated.

In this study, correlation pleiades are appropriate since the drugs can be grouped in clusters with highly correlated use (see Fig 4.10A and 4.10B):

- The *Heroin pleiad (heroinPl)* includes crack, cocaine, methadone, and heroin;
- The *Ecstasy pleiad (ecstasyPl)* includes amphetamines, cannabis, cocaine, ketamine, LSD, magic mushrooms, legal highs, and ecstasy;
- The *Benzodiazepines pleiad (benzoPl)* includes methadone, amphetamines, cocaine, and benzodiazepines.

These correlation pleiades include 12 drugs (Fig. 4.18). Additionally, we can consider the 'smoking couple', the highly correlated pair cannabis–nicotine. Other drugs do not have strong symmetric correlations. There exists an asymmetric correlation link from ecstasy to amyl nitrite (Fig. 4.11). Therefore, amyl nitrite can be considered as a peripheral element of the ecstasy pleiad.

Fuzzy and probabilistic clustering may help to reveal more sophisticated relationships between objects and clusters. For example, analysis of the intersections

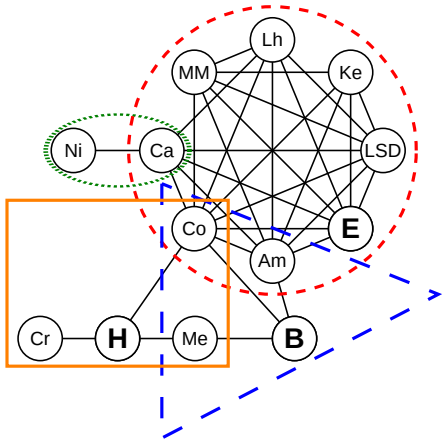

Figure 4.18. Correlation pleiades for drug use (in a circle, in a triangle and in a rectangle). Additionally, a highly correlated 'smoking couple', cannabis and nicotine, is separated by an ellipse. E stands for ecstasy, H for heroin, B for benzodiazepines, and MM for magic mushrooms. Other drugs are denoted by the first two letters of their names. Edges represent correlations.

between correlation pleiades of drugs can generate important question and hypotheses:

- Which patterns of behaviour are reflected by the existence of pleiades? (For example, is the ecstasyPl just the group of party drugs united by habits of use?)
- Why is cocaine a peripheral member of all pleiades?
- Why does methadone belong to the periphery of both the heroin and benzodiazepines pleiades?
- Why do amphetamines belong to the periphery of both the ecstasy and benzodiazepines pleiades?
- Do these intersections reflect the structure of individual drug consumption or the structure of the groups of drug consumers?

We define groups of users and non-users for each pleiad. *A group of users for a pleiad includes the users of any individual drug from the pleiad* (see Table 4.25). A group of non-users contains all participants which are not included in the group

Table 4.25. Number of drug users for pleiades in the database

Pleiad	User definition based on			
	Decade	Year	Month	Week
HeroinPl	832 (44.14%)	585 (31.03%)	309 (16.39%)	184 (9.76%)
EcstasyPl	1317 (69.87%)	1089 (57.77%)	921 (48.86%)	792 (42.02%)
BenzoPl	1089 (57.77%)	830 (44.03%)	528 (28.01%)	363 (19.26%)

of users. Table 4.25 shows the total number of users and their percentages in the database for three pleiades and for each user definition (the decade-, year-, month-, or week-based user/non-user separation).

The class imbalance problem is well known [117]. Users form a small fraction of the dataset (significantly less than half) for most of drugs (see Table 2.6). The classes of users and non-users are more balanced for pleiades of drugs than for individual drugs (compare Table 4.25 and 2.6). Table 4.25 shows that the number of drug users in the database for all three pleiades are more balanced (closer to 50%) than the number of users of the corresponding individual drug (Table 2.6). For example, for the decade-based classification problem the number of benzoPl users is 1089 (57.77%), while the number of benzodiazepines users is 769 (40.80%) and the number of heroinPl users is 832 (44.14%), while the number of heroin users is 212 (11.25%).

The introduction of moderate subcategories of T-$score_{sample}$ for pleiades of drugs enables us to separate the pleiades of drugs into two groups for the decade-, month-, and week-based user/non-user separation. For year-based user/non-user separation there is only one group with profile $(+, 0, +, -, -)$, and includes the users of heroinPl, ecstasyPl and benzoPl.

For the decade-based classification problem, the group with the profile $(+, 0, +, -, -)$ includes the users of heroinPl and benzoPl. The group with the profile $(0, 0, +, -, -)$ includes the users of EcstasyPl.

For the month- and week-based classification problem, the group with the profile $(+, -, +, -, -)$ includes the users of heroinPl and benzoPl. The group with the profile $(0, 0, +, -, -)$ includes the users of EcstasyPl.

Table 4.26. Statistically significant differences of means for groups of users and non-users for each pleiad for decade- year-, month-, and week-based classification problem. The symbol ' ⇓ ' corresponds to a significant difference where the mean in the users group is less than the mean in non-users group, and the symbol '⇑' corresponds to a significant difference where the mean in users group is greater than the mean in non-users group. Empty cells corresponds to insignificant differences. The difference is considered to be significant if the p-value is less than 0.01).

Pleiades of drugs	N	E	O	A	C
The decade-based user/non-user separation					
HeroinPl, EcstasyPl, BenzoPl	⇑		⇑	⇓	⇓
The year-based user/non-user separation					
HeroinPl, EcstasyPl, BenzoPl	⇑		⇑	⇓	⇓
The month-based user/non-user separation					
HeroinPl, EcstasyPl	⇑		⇑	⇓	⇓
BenzoPl	⇑	⇓	⇑	⇓	⇓
The week-based user/non-user separation					
HeroinPl, BenzoPl	⇑	⇓	⇑	⇓	⇓
EcstasyPl	⇑		⇑	⇓	⇓

The personality profiles for pleiades of drugs are qualitatively similar but some differences should be mentioned: the N level for EcstasyPl users is lower than for HeroinPl users, whereas levels of E and A are higher for EcstasyPl users (see Fig. 4.19, Fig. 4.20, Fig. 4.21 and 4.22).

We have applied the eight methods described in Section 'Risk evaluation methods' and selected the best one for each pleiad for the decade-, year-, month-, and week-based classification problems. The results of the classifier selection are pre-

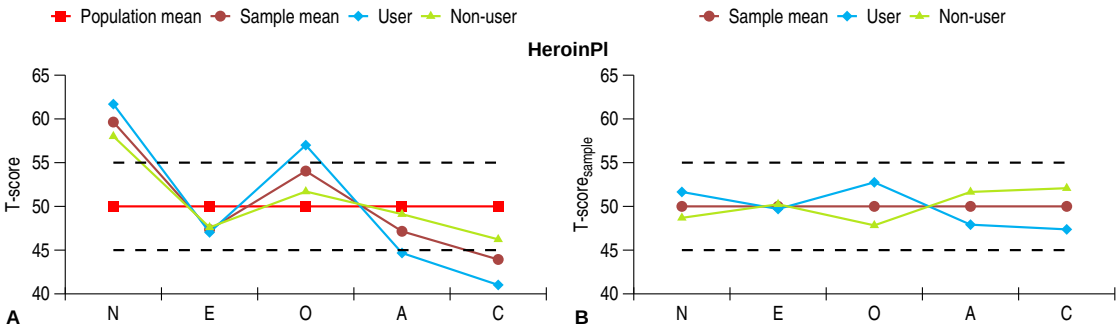

Figure 4.19. Average personality profiles for HeroinPl for the decade-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means.

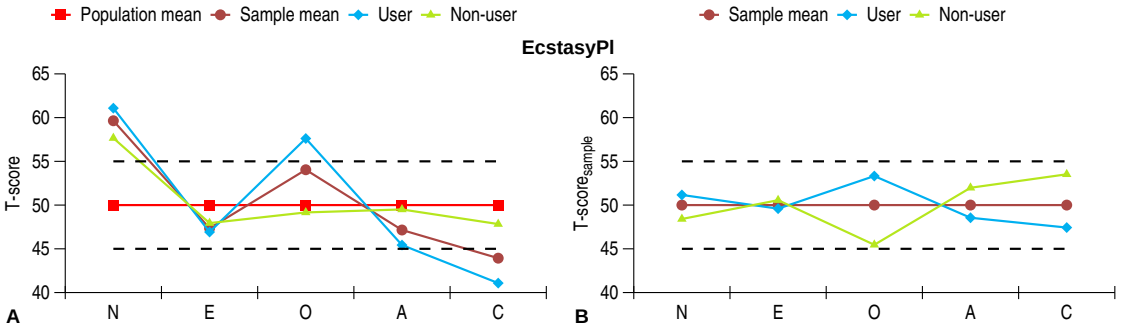

Figure 4.20. Average personality profiles for EcstasyPl for the month-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means.

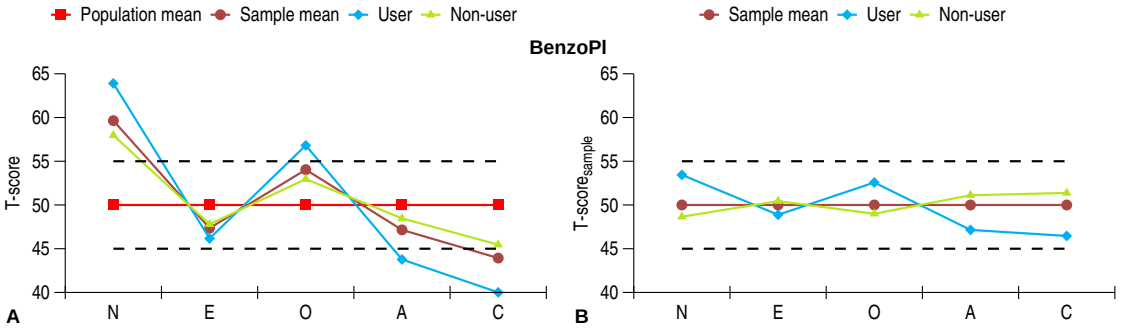

Figure 4.21. Average personality profiles for BenzoPl for the month-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means.

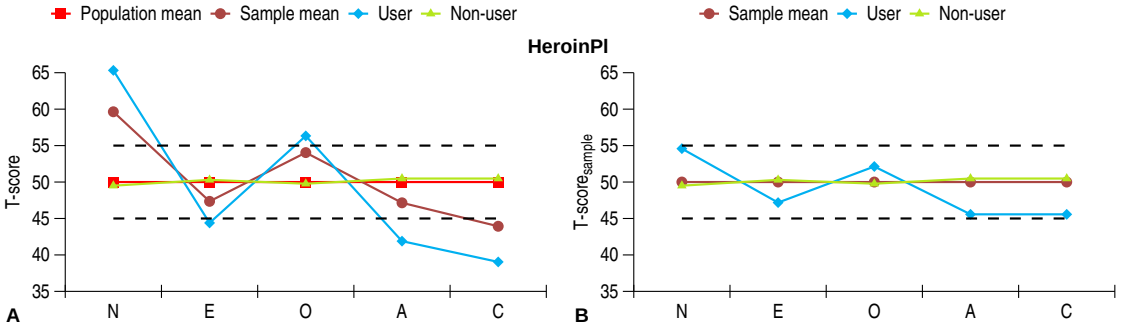

Figure 4.22. Average personality profiles for HeroinPl for the week-based user/non-user separation. A: T-scores with respect to the population norm mean and B: T-score$_{sample}$ with respect to the sample means.

Table 4.27. The best results of the pleiad users classifiers. Symbol 'X' means input feature used in the best classifier. Sensitivity and Specificity were calculated by LOOCV.

Pleiades of drugs	Classifier	Age	Edu	N	E	O	A	C	Imp	SS	Gndr	#	Sn (%)	Sp (%)	Sum (%)
The decade-based user/non-user separation															
HeroinPl	DT				X	X		X			X	4	71.23	78.85	150.07
EcstasyPl	DT	X		X	X	X				X	X	6	80.63	79.80	160.44
BenzoPl	DT		X		X	X	X			X		5	73.37	72.45	145.82
The year-based user/non-user separation															
HeroinPl	DT		X	X			X				X	4	73.69	71.80	145.49
EcstasyPl	DT	X	X			X		X	X	X	X	7	80.65	80.72	161.37
BenzoPl	DT		X	X				X		X		4	73.93	73.98	147.91
The month-based user/non-user separation															
HeroinPl	DT	X			X	X		X		X		5	74.18	74.11	148.29
EcstasyPl	PDFE	X	X		X	X		X	X	X	X	8	79.34	79.50	158.83
BenzoPl	DT	X	X				X	X				4	73.18	73.11	146.28
The week-based user/non-user separation															
HeroinPl	DT	X	X	X	X	X			X	X	X	8	75.84	73.91	149.75
EcstasyPl	LR	X	X		X	X	X	X		X	X	8	77.68	77.78	155.45
BenzoPl	DT		X				X	X		X	X	5	75.10	75.76	150.86

sented in Table 4.27. The quality of classification is high.

The classification results are excellent for each pleiad for the decade-, year-, month-, and week-based problems. We can compare the classifiers for one pleiad and for different problems (see Table 4.27). For example,

- The best classifier for ecstasyPl for the year-based user/non-user separation is a DT with seven attributes and has sensitivity 80.65% and specificity 80.72%.
- The best classifier for heroinPl for the month-based user/non-user separation is a DT with five attributes and has sensitivity 74.18% and specificity 74.11%.
- The best classifier for benzoPl for the week-based user/non-user separation is a DT with five attributes and has sensitivity 75.10% and specificity 75.76%.

Comparison of Tables 4.18 and 4.27 shows that the best classifiers for the ecstasy

and benzodiazepines pleiades are more accurate than the best classifiers for the consumption of the 'central' drugs of the pleiades, ecstasy and benzodiazepines respectively, even for the decade-based user definition. Classifiers for heroinPl may have slightly worse accuracy but these classifiers are more robust because they solve classification problems which have more balanced classes. All other classifiers for pleiades of drugs are more robust too for the same reasons, for all pleiades and definitions of users.

Tables 4.18 and 4.27 for the decade-based user definition show that most of the classifiers for pleiades use more input features than the classifiers for individual drugs. We can see from these tables that the accuracies of the classifiers for pleiades and for individual drugs do not differ drastically, but the use of a greater number of input features suggests more robust classifiers.

It is important to remark that pleiades are usually assumed to be disjoint. We consider pleiades which are named by the central drug and the peripheral drugs can be shared. For example, all three pleiades have cocaine as an intersection. This approach corresponds to the concept of 'soft clustering'.

4.11 Overoptimism problem

The selected best machine learning methods give impressive solutions of the user/non-user classification problems. Nevertheless, the procedure of selection has used the same data as the training process: we test each method by LOOCV. Such an approach could produce the so-called overoptimism: the cross-validation errors of the best method on the same set, which was used for the method selection, may be underestimated.

To prove that the data of tables 4.18 and 4.27 are valid for generalisation errors and the samples we have never seen before, we may need additional validation on large hold out sample, which was not compromised by using in the method selection. We do not have additional large sample and splitting the existing sample

into training set (for training and cross-validation in method selection) and for validation set (for validation of the best method) will decrease the *statistic power* of analysis [140].

Following [142], high performance on the test sample does not guarantee high performance on future samples, things do change and there is always a chance that a variable and its relationships will be different in the future samples. Selection of the best models and best sets of 'dominant variables' can damage the model robustness to the future variations.

The idea of stability of the model can significantly help in the classifiers testing [141]. In the process of cross-validation we can test additionally stability of the model and answer the questions:

- How many examples change their class in cross-validation (we can calculate the number for each transition between classes: class A→ class B, etc.). This is *classification stability*.
- How many qualitatively different models (for example, decision trees with different structure) were generated in cross-validation. This is *structural stability*.

We can also extract the set of examples with unstable classification and study this set separately.

Hand [143] clearly demonstrated that 'simple methods typically yield performance almost as good as more sophisticated methods, to the extent that the difference in performance may be swamped by other sources of uncertainty that generally are not considered in the classical supervised classification paradigm.'

Therefore, let us consider the results of the best methods (tables 4.18 and 4.27) as a upper border of the possible classifier performance and apply the simple method, linear discriminant. This method is robust and leaves no space for overoptimism if the samples are sufficiently large and there is no multicollinearity. We will also analyse the classification stability of the linear discriminant in cross-validation.

Multicollinearity means strong linear dependence between attributes. It makes the model very sensitive to fluctuations in data and can be considered as an important source of instability of classifiers . The tests of multicollinearity are based on the analysis of efficient and stable invertibility of the correlation matrix [144]. One of the standard measures of multicollinearity is the condition number of the correlation matrix, that is the ratio $\kappa = \lambda_{\max}/\lambda_{\min}$, where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximal and the minimal eigenvalues of this matrix. (Let us recall that the correlation matrix is symmetric and non-negative and its eigenvalues are real and non-negative numbers.) The collinearity with $\kappa < 10$ is considerated as 'modest' collinearity and most of methods work in this situation and only few methods were reported as failed (like support vector machine) [145].

Eigenvalues of correlation matrix between seven psychological traits are:

$$2.267, 1.809, 0.887, 0.678, 0.548, 0.468, 0.342;\ \kappa = 6.628.$$

Eigenvalues of correlation matrix between ten attributes (quantified) including seven psychological traits, Age, Edu and Gndr are:

$$2.595, 1.867, 1.111, 0.980, 0.814, 0.757, 0.599, 0.524, 0.427, 0.327;\ \kappa = 7.945.$$

We can see that there is no strong multicollinearity despite of existence of significant and not small correlations between psychological traits (see tables 4.3 and 4.4). Three correlation coefficients: between the N and E scores, between the O and SS scores and between the Imp and SS scores exceed 0.4 by absolute value. Correlation absolute values above 0.4 are sometimes interpreted as indicating a multicollinearity problem. This heuristic rule is not rigorous but existence of such correlations rises a question about multicollinearity and it is necessary to apply a stronger test. We calculated the conditional number and it is sufficiently low to exclude strong multicollinearity. In the next section we deminstrate that Fisher's linear discriminant is sufficiently robust and works stable for these values of κ in

this database.

4.12 User/non-user classification by linear discriminant for ecstasy and heroin

Fisher [134] defined the linear discriminant (LD) as a linear function of attributes, for which the ratio of the difference between classes to the standard deviation within classes is maximal (see, for example, the score 4.1.1). We used the classical formula for the LD direction (3.3.2). The separation threshold (intercept) Θ is defined by the balance condition Sn=Sp.

Linear discriminants separate users from non-users by linear inequalities:

$$D(z) = \Theta + \sum c_i z_i > 0 \tag{4.12.1}$$

for users and ≤ 0 for non-users, where Θ are the thresholds (intercepts), z_i are the attributes, and c_i are the coefficients.

In the Appendix Tables C.7-C.10 contain the coefficients c_i of linear discriminants for user/nonuser separation in 10-dimensional space (7 psychological attributes, Age, Edu, and Gndr). The attributes in these tables are quantified and transformed to z-scores with zero mean and unit variance (positive values of the Gndr z-score corresponds to female). The last rows of the tables include the standard deviation of the coefficients in LOOCV. For 7-dimensional space of psychological attributes (T-scores), the coefficients of linear discriminants are presented in tables C.15-C.18.

Performance of linear discriminants in user/non-user separation is evaluated by several methods (tables C.11–C.14 for 10-dimensional data space and tables C.19–C.22 for 7-dimensional space of T-scores of psychological attributes). First of all, we calculated the linear discriminant using the whole sample (see tables C.7–C.10) and find all their errors. For each solution of the classification problem we

Table 4.28. Coefficients of linear discriminant for ecstasy user/non-user separation and decade-, year-, month-, and week-based definition of users (10 attributes)

Period	Θ	Age	Edu	N	E	O	A	C	Imp	SS	Gndr
Decade	-0.171	-0.631	-0.188	0.053	0.018	0.351	-0.065	-0.210	-0.088	0.559	-0.265
Year	-0.464	-0.782	-0.101	-0.015	0.099	0.238	-0.025	-0.173	-0.004	0.453	-0.275
Month	-0.633	-0.820	0.047	-0.139	0.093	0.284	-0.123	-0.165	-0.028	0.328	-0.257
Week	-0.779	-0.697	0.077	-0.115	0.161	0.545	-0.093	-0.252	0.217	0.230	-0.022
SD	≤ 0.018	≤ 0.002	≤ 0.003	≤ 0.004	≤ 0.004	≤ 0.003	≤ 0.003	≤ 0.003	≤ 0.004	≤ 0.005	≤ 0.003

have several numbers, P (positive), the number of samples recognised as positive, and N - negative, the number of samples recognised as negative. $P + N$ is the total number of samples. P=TP+FP (True Positive plus False Positive) and N=TN+FN (True Negative plus False Negative). Sensitivity is Sn=TP/(TP+FN)×100% and Specificity is Sp=TN/(TN+FP)×100%. Accuracy is Acc=(TP+TN)/(P+N)×100%.

We calculate these performance indicators for the total sample and for the LOOCV procedure. In LOOCV the linear discriminant is calculated for the set of all samples excluding the example left out for testing. The test was performed for all samples with the corresponding redefining of Sn, Sp, and Acc. In LOOCV the linear discriminants are calculated for each testing example. Each of these discriminants is a separate classification model. Stability of classification can be measured by the number of examples which change their class at least once. We took the basis model for the total sample and find how many true positive examples of this model became FN examples of a LOOCV model at least once. This number measured in % of TP+FP of the basic model is TP→FN. Analogously, we defined FP→TN, TN→FP, and FN→TP. The last two numbers are measured in % of TN+FN of the basic model.

In this Section we analyse performance of linear discriminants for two drugs, ecstasy and heroin. They are typical (and central) elements of two pleiades, groups of drugs with correlated drug usage. The differences between them could tell us a story about different types of drug users.

Coefficients of LD can be used for indication how a change in an attribute value will affect the value of D(z) under condition that the values of all other attributes do not change. It is possible to change one attribute without changing other at-

Table 4.29. Performance and stability of linear discriminant for ecstasy user/non-user separation and decade-, year-, month-, and week-based definition of users (10 attributes). All indicators are in %.

Period	Total sample			LOOCV			Stability indicators			
	Sn	Sp	Acc	Sn	Sp	Acc	TP→FN	FN→TP	FP→TN	TN→FP
Decade	74.4	74.7	74.6	74.2	74.3	74.2	0.3	0.7	0.5	0.4
Year	75.4	75.7	75.6	75.0	75.6	75.4	0.2	0.6	0.8	0.3
Month	72.5	72.4	72.4	71.3	72.3	72.1	1.7	1.3	1.2	1.2
Week	72.6	71.5	71.5	67.9	71.4	71.2	3.6	8.3	2.4	5.3

Table 4.30. Coefficients of linear discriminant for heroin user/non-user separation and decade-, year-, month-, and week-based definition of users (10 attributes)

Period	Θ	Age	Edu	N	E	O	A	C	Imp	SS	Gndr
Decade	-0.615	-0.210	-0.370	0.413	-0.211	0.477	-0.265	-0.029	0.222	0.381	-0.332
Year	-0.849	-0.584	-0.168	0.352	-0.252	0.222	-0.275	0.014	0.216	0.359	-0.378
Month	-1.037	-0.560	-0.371	0.181	-0.350	0.159	-0.397	0.016	0.368	0.154	-0.226
Week	-1.096	-0.386	-0.077	0.467	-0.255	0.184	-0.412	0.013	0.437	-0.077	-0.400
SD	≤ 0.005	≤ 0.004	≤ 0.004	≤ 0.004	≤ 0.004	≤ 0.004	≤ 0.004	≤ 0.004	≤ 0.005	≤ 0.006	≤ 0.003

Table 4.31. Performance and stability of linear discriminant for heroin user/non-user separation and decade-, year-, month-, and week-based definition of users (10 attributes). All indicators are in %.

Period	Total sample			LOOCV			Stability indicators			
	Sn	Sp	Acc	Sn	Sp	Acc	TP→FN	FN→TP	FP→TN	TN→FP
Decade	70.8	69.9	70.0	68.4	69.8	69.7	2.8	6.1	1.3	2.2
Year	73.7	73.7	73.7	70.3	73.6	73.4	2.5	2.5	2.1	2.3
Month	79.2	77.6	77.7	69.8	77.5	77.2	9.4	7.5	3.2	3.8
Week	79.3	80.1	80.1	65.5	80.0	79.8	6.9	6.9	4.6	4.8

tributes because there is no strong multicollinearity. The most interesting in this ranking for the tables 4.28 and 4.30 is the essential difference between ranking for ecstasy and for heroin user/non-user discriminants.

Comparison of tables 4.28 and 4.30 gives immediately a result: the coefficients of LD for ecstasy and heroin have significant differences: for ecstasy, effect of Imp is less than for heroin (and can even have different sign), whereas effect of SS is bigger for ecstasy. For ecstasy, the coefficients at A have smaller values than the coefficients at C. For heroin the situation is inverse: C has much less coefficient than A (and can even have the opposite sign). Also, coefficients at N for ecstasy are smaller than for heroin and can have different sign. Edu has for heroin negative coefficients with bigger values (it 'prevents' usage of heroin), whereas for ecstasy influence of Edu is smaller and can have opposite signs (for week- and month-

based definition of usage). Coefficient at E are positive for ecstasy user/non-users separation and negative for heroin user/non-user separation. Age has large negative coefficients both for ecstasy and heroin but for ecstasy they are 1.5–2 times bigger. For example, for month-based definition of users the ranking of variables is:

- For ecstasy: Age⇓ , SS⇑, O⇑, Gndr⇓ , C⇓, N⇓, A⇓, E⇑, Edu⇑, Imp⇓;
- For heroin: Age⇓, A⇓, Edu⇓, Imp⇑, E⇓, Gndr⇓ , N⇑, O⇑, SS⇑, C⇑.

The arrows ⇑ ⇓ here indicates the sign of the effect of the attribute in the user/non-user separation (and not the shift of the mean as it was in previous sections): for positive coefficient it is ⇑ and for negative coefficients it is ⇓. The difference between heroin and ecstasy discriminants is impressive. The most important five attributes for ecstasy user/non-user discrimination have only one attribute in common with the top five attribute of heroin user/non-user discrimination (Age). There are several attributes with different signs of coefficients for ecstasy and heroin linear user/non-user discrimination in this case: C, N, E, and Imp For most of these attributes, the traditional expectation is well-known: C⇓, N⇑, Imp⇑ (at least, for illegal drugs). The values of the coefficients with unexpected signs are relatively small, the attributes with these values are ranked as less important, but the expectation is not met in any case: we can state that it is wrong assumption that high N and Imp are predictors for ecstasy use and it is also wrong that low C is predictor for heroin use.

If we do not use Age, Edu, and Gndr, then the difference of the predictors persists (tables 4.32 and 4.33): for ecstasy LD the most important attribute becomes SS, then O and C. The attributes Imp, A, N, and E have smaller coefficients, and for Imp, E, and N the sign of the coefficients depends on recency of usage. For heroin C seems to be less important than other attributes.

The ranking of these seven psychological attributes for the same month-based user/non-user discrimination looks similarly:

Table 4.32. Coefficients of linear discriminant for ecstasy user/non-user separation and decade-, year-, month-, and week-based definition of users. (7 attributes)

Period	Θ	N	E	O	A	C	Imp	SS
Decade	-35.896	0.045	-0.017	0.407	-0.085	-0.342	-0.156	0.827
Year	-42.579	-0.019	0.078	0.373	-0.045	-0.356	-0.096	0.846
Month	-27.572	-0.169	0.101	0.462	-0.191	-0.354	-0.139	0.752
Week	-60.127	-0.095	0.168	0.695	-0.128	-0.355	0.242	0.528
SD	≤ 0.565	≤ 0.004	≤ 0.005	≤ 0.004	≤ 0.004	≤ 0.004	≤ 0.005	≤ 0.006

Table 4.33. Coefficients of linear discriminant for heroin user/non-user separation and decade-, year-, month-, and week-based definition of users. (7 attributes)

Period	Θ	N	E	O	A	C	Imp	SS
Decade	-55.851	0.381	-0.283	0.526	-0.322	-0.124	0.249	0.563
Year	-44.733	0.361	-0.327	0.368	-0.397	-0.108	0.218	0.641
Month	-30.020	0.302	-0.409	0.232	-0.499	-0.107	0.345	0.555
Week	-37.099	0.548	-0.325	0.265	-0.547	-0.069	0.389	0.261
SD	≤ 0.523	≤ 0.004	≤ 0.004	≤ 0.005	≤ 0.004	≤ 0.005	≤ 0.006	≤ 0.007

- For ecstasy: SS⇑, O⇑, C⇓, N⇓, A⇓, Imp⇓, E⇑;
- For heroin: SS⇑, A⇓ , E⇓ , Imp⇑, N⇑, O⇑, C⇓ .

We have to stress the opposite signs at N, E, and Imp for the ecstasy and heroin user/non-user discriminants in this case. The only big jump from the 10-attribute ranking is the change of rank of SS for heroin user/non-user discrimination. We can guess that this is because of large negative correlations between SS and Age, which is important for classification but not available in the seven-attribute model. Again, the upper four variables for ecstasy user/non-user discrimination have only one attribute in common with the top four attribute of heroin user/non-user discrimination (SS).

Already simple LDA demonstrates that users of ecstasy and heroin differ significantly and users of different groups of drugs should be studied separately. Just the hypothesis that drug usage is associated with N⇑, A⇓, and C⇓ seems plausible from the first glance, but appears to be oversimplification. Such an analysis and even more detailed consideration of all definitions of drug use is possible for every pair of drugs (and for four groups of drugs) on the basis of the linear discriminant coefficients presented in tables C.7-C.10 and C.15-C.18.

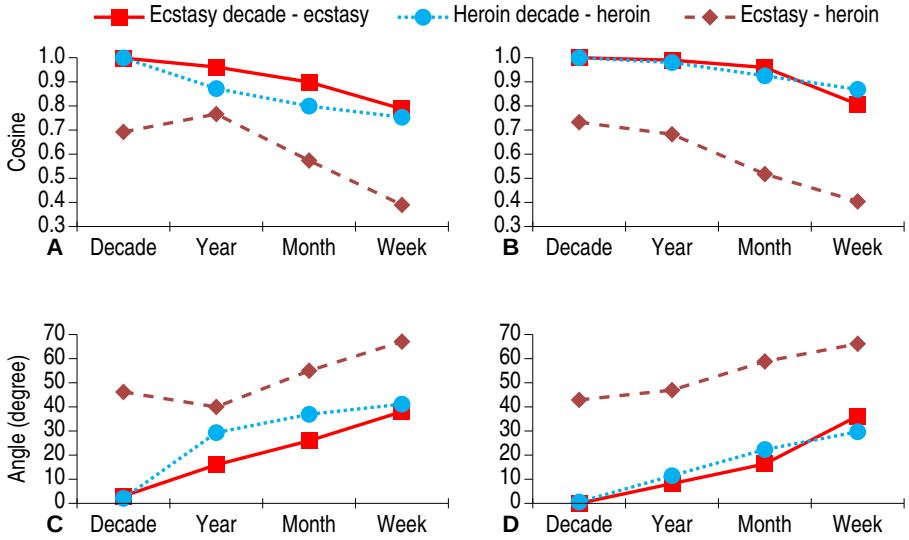

Figure 4.23. Angles between directions of linear discriminants for user/non-user classification for ecstasy and heroin. Two types of angels are presented: between the discriminant directions for all periods and the discriminant vector for decade-based definitions of users for both drugs and angles between directions of linear discriminants for ecstasy and heroin (and the same periods). For convenience, both cosines of angles (A, B) and angles in grads (C, D) are presented.

The difference between linear discriminant directions can be evaluated by angles between them. We calculate the angles between linear discriminant directions for all definition of users and this direction for the decade-based definition of users (for the same drug). These angles are significantly smaller than the angles between the linear discriminant directions for the same recency of use and different drugs (Fig. 4.23).

4.13 Separation of heroin users from ecstasy users

In this Section we continue study of difference between heroin and ecstasy use. The first simple question is: what is the intersections between the sets of users? The answer is illustrated by Fig. 4.24. It is obvious from the figure that for recent users the proportion of simultaneous use of heroin and ecstasy becomes smaller. We can hypothesize that the people who used drugs more than month or year ago but not recently just tried various drugs, whereas recent users prefer more

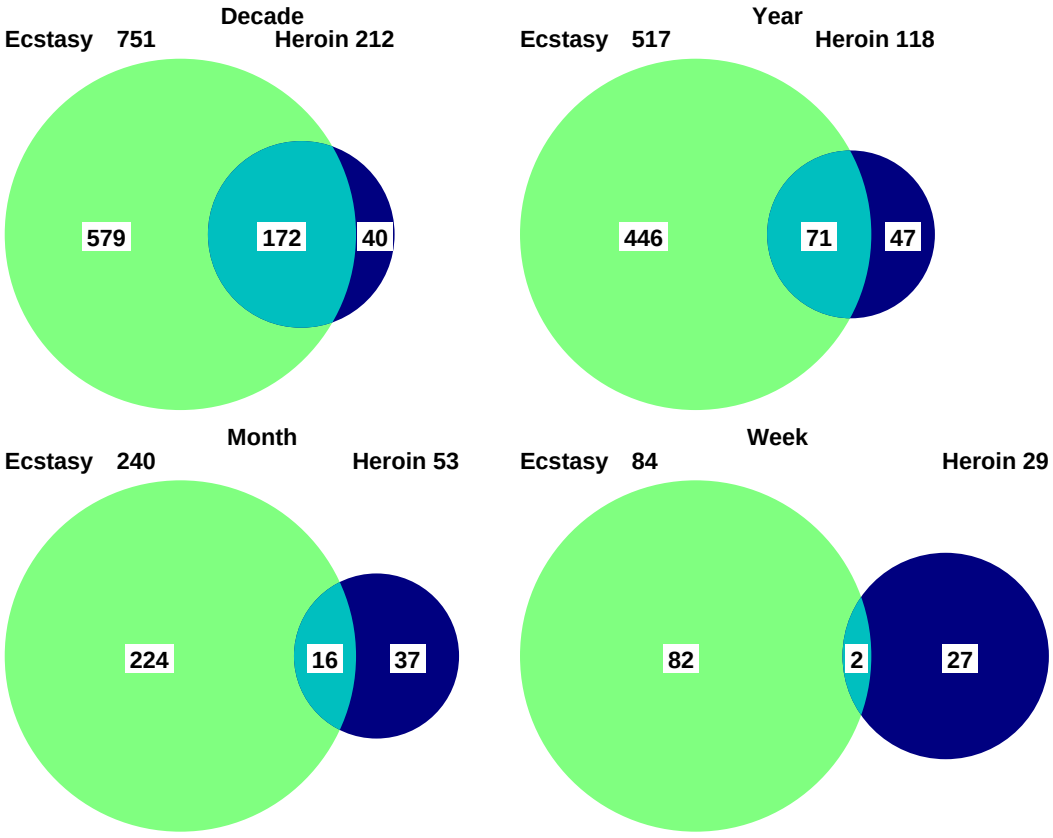

Figure 4.24. Venn diagrams of relations between ecstasy and heroin use for decade-, year-, month-, and week-based definitions of users.

specific drugs. Fig. 4.24 is an argument in favor of this hypothesis.

For each recency of use, there are six important sets: users of ecstasy, users of heroin, users of ecstasy OR heroin (the union), users of ecstasy AND heroin (the intersection), users of ecstasy NOT heroin (users of ecstasy only, the difference: users of ecstasy \ users of heroin), and users of heroin NOT ecstasy (users of heroin only, the difference: users of heroin \ users of ecstasy).

The intersection of heroin and ecstasy users is large for decade-based user definition and decreases for more recent users (Fig. 4.24). The discrimination of ecstasy and heroin users is a non-standard classification problem because this intersection. At least, the standard TPR (Sn) and TNR (Sp) have not much sense. Let us consider a binary classification rule which separates all users of ecstasy OR heroin into two classes: E and H. We will consider an example from the set of users of heroin OR ecstasy as FE ('false ecstasy') if it is a non-user of ecstasy classified as

a user of ecstasy (or, which is the same, a user of heroin but NOT ecstasy classified as a user of ecstasy). Analogously and example is considered as FH ('false heroin') if it is non-user of heroin classified as a user of heroin. In the tables 4.34 we use and unusual measures of classification accuracy: True Ecstasy Rate (TER) (correctly recognised fraction of users of ecstasy NOT heroin) and True Heroin Rate (THR) (correctly recognised fraction of users of heroin NOT ecstasy). We do not consider as an error a case when a user of both drugs is recognised as a user of one them:

$$\text{TER} = \frac{\text{\# correctly recognised users of ecstasy NOT heroin}}{\text{\# users of ecstasy NOT heroin}};$$

$$\text{THR} = \frac{\text{\# correctly recognised users of heroin NOT ecstasy}}{\text{\# users of heroin NOT ecstasy}}.$$

The descriptive statistics for seven traits (five factor model, Imp, and SS) are presented in Tables 4.34 (compare to table 4.1).

We can see that for ecstasy - heroin discrimination the best classifier with one attribute is N. Moreover, differences between means of ecstasy users and heroin users are statistically significant with confidence level 99% for N and A for all definitions of users, for Imp for decade-based and month-based definitions and for E for all definitions excluding decade-based one. This difference is obvious in the graphs of mean values of N, E, A, and Imp for ecstasy and heroin users presented in Fig. 4.26.

Let us employ LDA for separation of ecstasy users from heroin users. We apply the formula for linear discriminant direction with covariance matrices for ecstasy users and heroin users calculated for month-based definition of users. The intercept Θ was calculated for the most balanced separation measured by TER and THR.

The ranking of the attributes for linear discriminant ecstasy users / heroin users

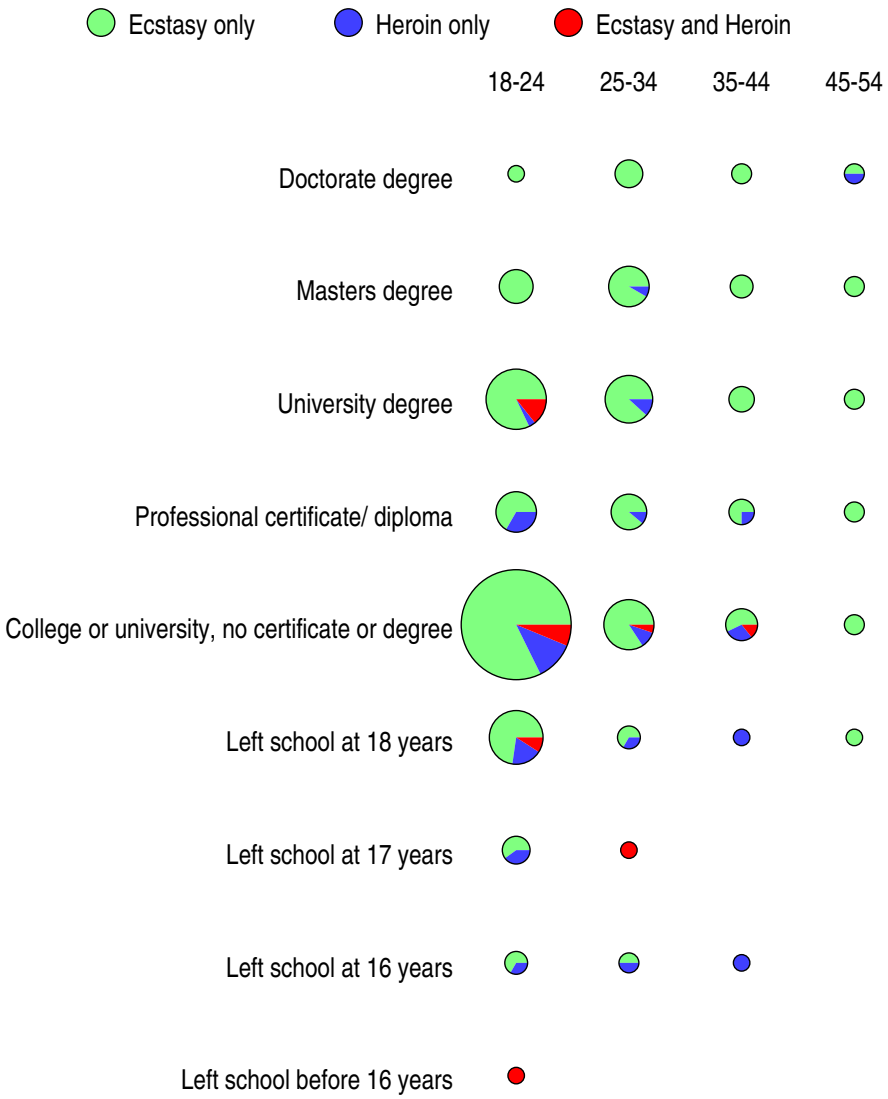

Figure 4.25. Distribution of ecstasy (NOT heroin) users, heroin (NOT ecstasy) users, and heroin AND ecstasy users in various age and education groups.

Table 4.34. Means and standard deviations for users of ecstasy, for users of heroin, and for users of ecstasy OR heroin. Dimensionless z-score (4.1.1) for separation of ecstasy users from heroin users is presented as well as $P = \phi(z)$, TER, and THR.

Factors	Ecstasy users		Heroin users		Users of one of them		H-E	One feature classifier				
	Mean	SD	Mean	SD	Mean	SD	p-value	z	P(%)	Θ	TER(%)	THR(%)
Decade-based definition												
N	25.06	9.20	28.06	8.89	25.26	9.23	< 0.001	0.154	56	27	65	61
E	27.91	7.12	26.48	7.37	27.76	7.14	0.012	0.088	54	27	57	60
O	36.14	6.02	36.55	5.86	36.07	6.02	0.366	0.041	52	35	55	55
A	29.88	6.61	27.98	7.28	29.77	6.67	0.001	0.128	55	28	62	58
C	27.50	6.87	26.52	7.00	27.45	6.87	0.071	0.067	53	27	53	55
Imp	4.45	2.06	5.02	2.06	4.44	2.06	< 0.001	0.139	56	4	48	58
SS	6.90	2.23	7.18	2.34	6.86	2.26	0.117	0.070	53	6	59	50
Year-based definition												
N	24.57	9.46	28.83	8.76	25.02	9.47	< 0.001	0.209	58	27	64	62
E	28.60	7.12	26.01	7.46	28.20	7.29	0.001	0.148	56	27	62	66
O	36.46	6.00	36.03	5.97	36.46	5.98	0.483	0.035	51	36	54	53
A	29.93	6.74	27.25	7.49	29.68	6.86	< 0.001	0.170	57	28	61	57
C	27.36	6.88	26.27	7.38	27.26	6.91	0.145	0.069	53	26	54	55
Imp	4.59	2.04	5.08	2.01	4.62	2.05	0.020	0.113	54	4	55	51
SS	7.14	2.15	7.29	2.36	7.15	2.18	0.525	0.031	51	7	55	50
Month-based definition												
N	23.50	9.73	30.04	8.76	24.32	9.80	< 0.001	0.308	62	27	68	66
E	29.09	7.53	24.58	8.65	28.43	7.89	0.001	0.233	59	27	64	65
O	36.66	6.12	35.40	6.82	36.55	6.14	0.217	0.089	54	36	55	54
A	29.64	6.82	25.83	7.42	29.27	6.81	0.001	0.242	60	28	59	57
C	27.54	7.20	24.81	7.60	27.19	7.31	0.020	0.159	56	26	58	59
Imp	4.53	2.07	5.30	1.89	4.63	2.07	0.010	0.170	57	4	65	52
SS	7.08	2.14	7.53	2.31	7.12	2.18	0.198	0.092	54	7	59	52
Week-based definition												
N	24.18	9.66	31.83	6.76	26.12	9.67	< 0.001	0.348	64	28	67	68
E	29.86	8.29	24.03	8.84	28.25	8.81	0.003	0.239	59	25	73	70
O	37.81	6.06	35.34	7.62	37.12	6.55	0.122	0.125	55	36	55	59
A	29.94	6.48	25.38	5.46	28.81	6.57	0.001	0.285	61	27	62	59
C	27.33	7.07	24.59	7.49	26.57	7.22	0.091	0.135	55	25	59	56
Imp	5.01	2.17	5.24	1.79	5.06	2.06	0.576	0.046	52	5	46	56
SS	7.33	2.19	7.28	2.45	7.30	2.25	0.911	0.005	50	7	57	48

Table 4.35. Coefficients (Coeff.) of linear discriminant for separation of ecstasy users from heroin users for month-based definition of users (7 attributes). TER=70.5% THR=73.0% (the sample of ecstasy AND heroin users); TER=69.6% THR=62.2% (LOOCV).

	Θ	N	E	O	A	C	Imp	SS
Coeff.	29.896	-0.541	0.476	-0.050	0.469	-0.136	-0.441	-0.214
SD	1.331	0.012	0.010	0.013	0.008	0.011	0.015	0.018

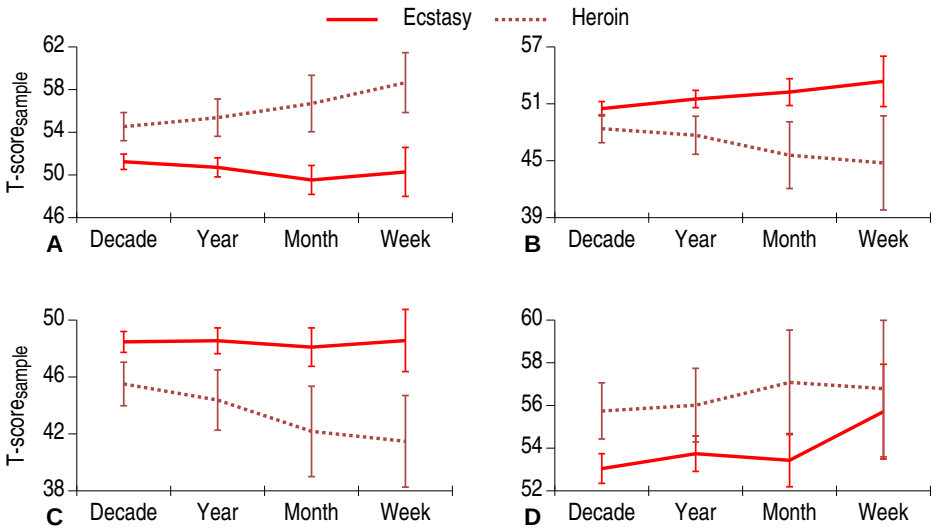

Figure 4.26. Mean values with their 95% confidence intervals for significantly different psychological traits of ecstasy and heroin users: A) N, B) E, C) A, D) Imp

Table 4.36. Coefficients (Coeff.) of linear discriminant for separation of ecstasy users from heroin users for month-based definition of users (10 attributes). TER=75.0% THR=73.0% (the sample of ecstasy AND heroin users); TER=71.6% THR=64.9% (LOOCV).

	Θ	Age	Edu	N	E	O	A	C	Imp	SS	Gndr
Coeff.	0.915	0.011	0.534	-0.401	0.379	-0.039	0.411	-0.176	-0.417	-0.092	0.166
SD	0.023	0.015	0.010	0.011	0.010	0.010	0.008	0.009	0.012	0.014	0.011

separation is (for month-based definition of users):

N, E, A, Imp, SS, C, O (7 attributes).

It is important to notice that the values of coefficients at N, E, A, and Imp do not differ much and for SS, C, and O the coefficients drop down fast.

For 10 attributes the ranking by the linear discriminant coefficients is:

Edu, Imp, A, N, E, C, Gndr, SS, O, Age.

Again, for the leading group of attributes, Edu, Imp, A, N, and E, the coefficients decay slowly and then drop down fast.

These observations together with analysis of attributes means (Fig. 4.26) ensures us that the main (and statistically significant) differences between ecstasy users and heroin users are in Edu, Imp, A, N, and E.

4.14 A tree of linear discriminants

Performance of simple LDA is not much worse than the quality of the best selected classifiers (see table 4.18). Moreover, the simple attempts to improve the performance by creating of a simple decision tree or simple kNN classifiers fail. In Fig. 4.27 a two-level decision tree with hierarchy of linear discriminants is presented. The PCA visualisation demonstrates that after the first linear discriminant application there remain a typical 'flies-and-mosquito' mixture without apparent user/non-user separation in the groups. (Principal components were recalculated for each node.) The kNN classifiers also do not demonstrate essential improvement of LDA (table 4.38).

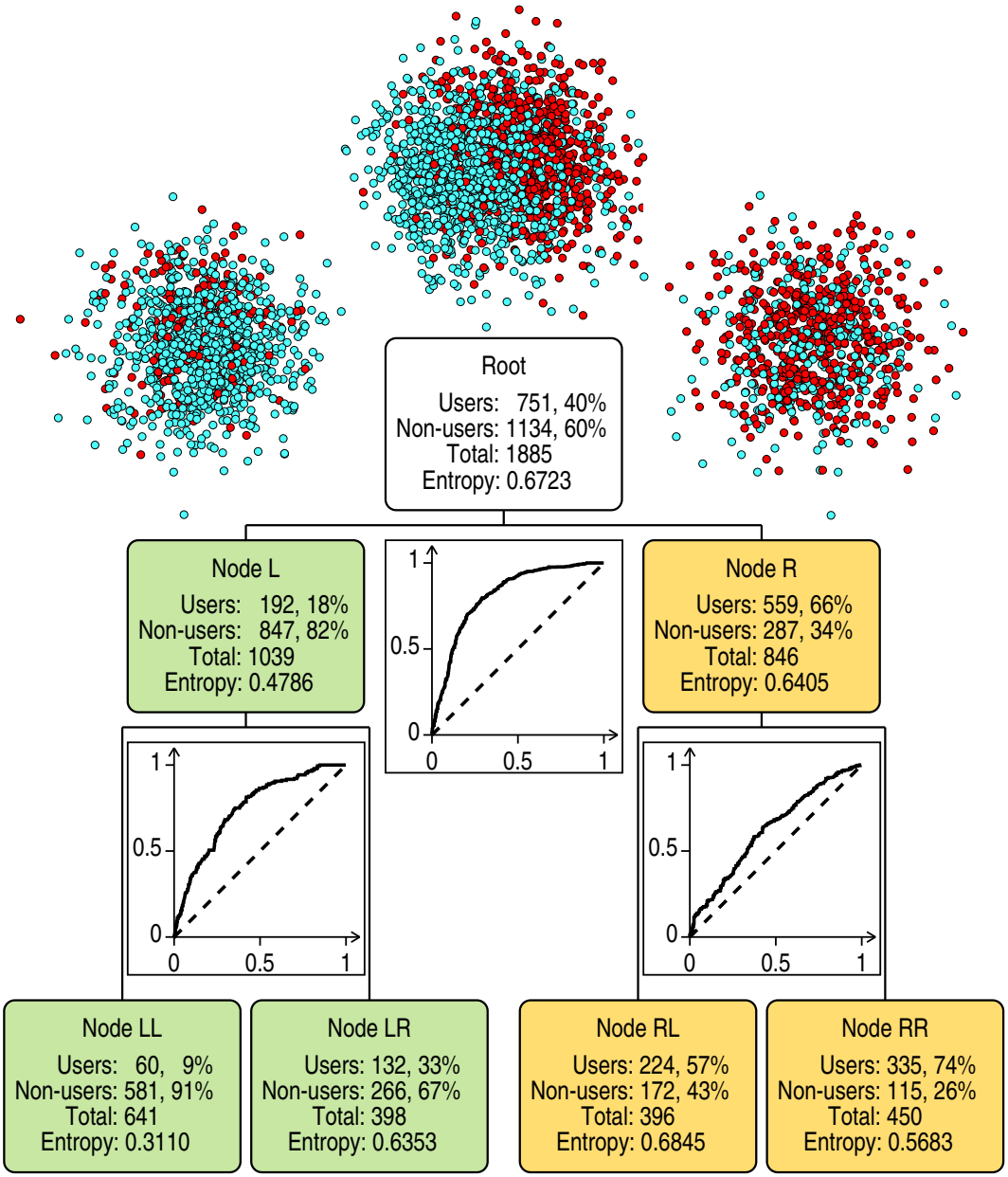

Figure 4.27. A two-level classification tree for ecstasy users and non-users (decade-based definition of users) with linear discriminant classifiers at the nodes. A data cloud is visualised by projection on the plain of two first principal components for the root and the nodes of the first levels (above nodes). Users are represented by blue (light) circles, non-users by red (dark) circles. The ROC curves for the linear discriminants at each branching node are below the nodes.

Table 4.37. The numbers of false positive (FP) and false negative (FN) errors for ecstasy user/non-user decision tree classifiers (decade-based definition of users) with linear discriminant at each node and with four different criteria of threshold selection: Accuracy, Sp+Sn, Balance (Sn=Sp), and Information Gain (IG).

	Accuracy		Sp+Sn		Balanced		IG	
Level	FP	FN	FP	FN	FP	FN	FP	FN
1	232	224	330	155	287	192	401	130
2	232	224	330	155	287	192	401	130
3	232	224	210	250	189	287	251	220
4	238	203	268	190	268	179	251	220
5	228	169	218	206	189	240	210	253
6	173	177	203	160	177	171	242	201
7	156	185	121	186	168	173	193	198

Table 4.38. Performance of kNN user/non-user classifiers for ecstasy (decade-based definition of users) for different k and for the standard Euclidean distance.

k	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Sp (%)	59.7	70.5	65.5	72.0	68.3	65.2	70.5	66.9	71.2	68.9	66.4	70.2	67.5	70.4	68.6
Sn (%)	81.9	72.8	80.2	72.2	78.3	81.5	75.8	78.8	74.7	77.8	80.2	77.5	80.0	76.0	79.5

4.15 Visualisation on non-linear PCA screen

Principal components provide us by a screen for convenient visualisation of data distribution (see Fig. 4.27). It works well if the data are distributed near a low-dimensional plane. Manifold learning methods of non-linear PCA generalise this idea to data approximation by smooth non-linear manifolds of small dimension [91]. These methods allow us to approximate data better and find in non-linear two-dimensional visual data maps the effects, which can be captured in higher-dimensional linear principal components only [90].

In a series of works the metaphor of elastic membrane and plate was used to construct one-, two- and three-dimensional principal manifold approximations of various topologies [90]. Mean squared distance approximation error combined with the elastic energy of the membrane serves as a functional to be optimised. The elastic map algorithm is extremely fast at the optimisation step due to the simplest form of the smoothness penalty.

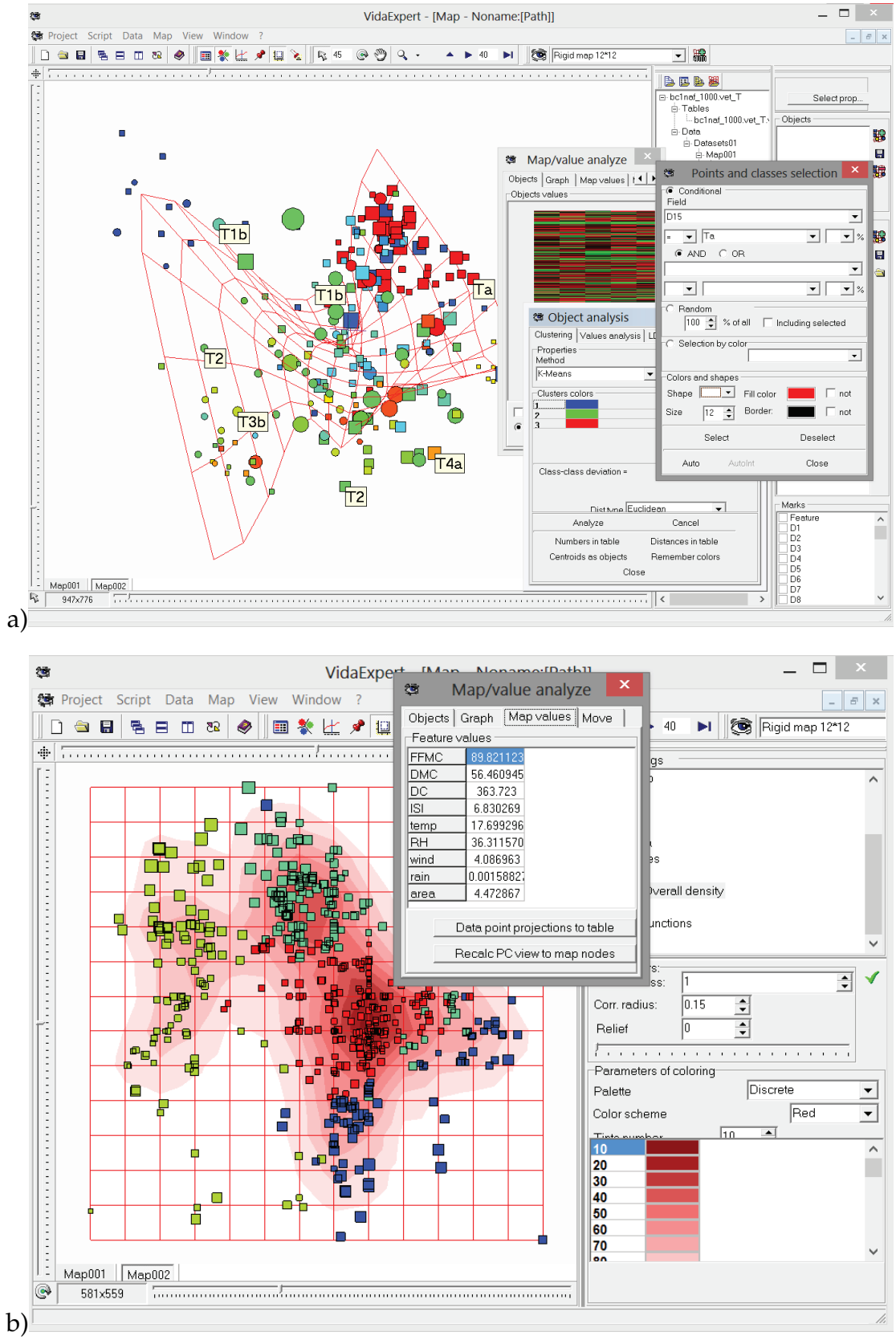

Figure 4.28. Srceenshots of VidaExpert: a) Elastic map in the three-dimensional PCA view, b) Coloring of the map in internal coordinates.

.

In this section we employed original software libraries ViDaExpert freely available online [146] (Fig. 4.28). This software allows to create an appropriate elastic manifold embedded in the dataspace (Fig. 4.28 a) and to color this map to visualise density and all the attributes (Fig. 4.28 b).

We created an elastic map for the whole dataset in the 10-dimensional space of quantified attributes (Fig. 4.30). Both three-dimensional PCA view and the elastic map view do not reveal any significantly non-linear non-ellipsoidal peculiarities in data distribution. In Figs 4.30 – 4.32 the attributes are visualised.

We can see that the attributes Imp, SS, and O generate similar colorings, which are opposite to coloring for age. It should be mentioned that this similarity means strong correlations of attributes on the two-dimensional map but does not imply the strong correlations in the higher-dimensional data space. Analogously, attributes N, A, C, and Edu have similar coloring on the map with N opposite to other three attribute colorings (Fig 4.31). The colorings for E and Gndr differ from all other (and are independent on the map) (Fig. 4.32).

Linear discriminant separation of drug users from non-users is visualised on the elastic map in Fig. 4.33. Of course, on the non-linear screen the linear discriminant is represented by a curve, not a straight line. The quality of the linear discriminant separation is visibly high.

4.16 Risk maps

For every set of attributes, we can evaluate the conditional probability density to be a drug user at each value of attributes. Visualisation of the conditional density of drug use can be considered as a risk map [147, 148]. Let us recall that the probability to be a drug user in the data base is higher than in the populations. In application of the risk maps to the real cases the risk evaluation should be renormalised to the population a priory probability of drug use. These maps can be used without such a renormalisation if we consider not the absolute risk but

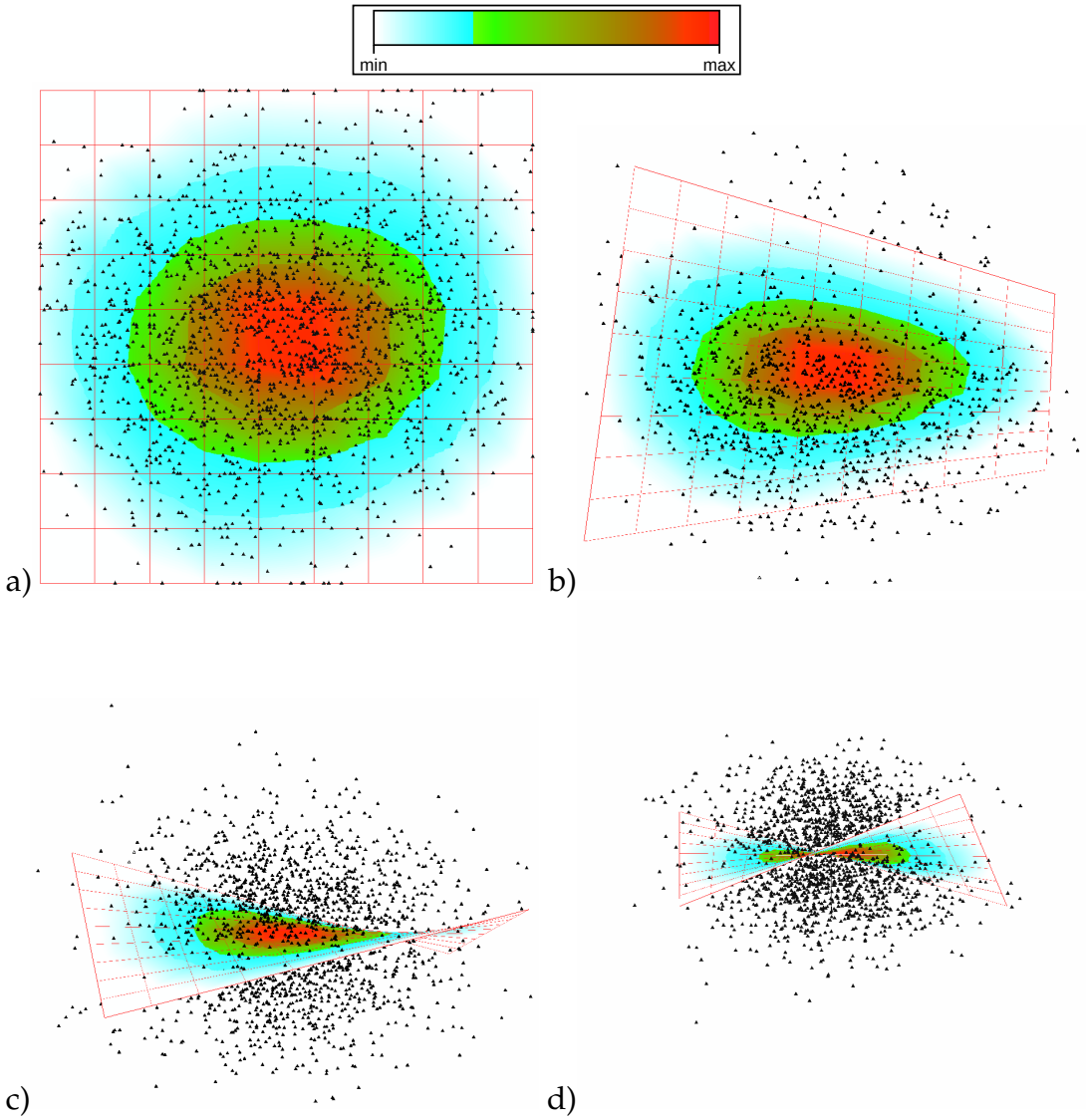

Figure 4.29. Elastic maps and density visualisation for the database of drug users: a) Density of the data cloud visualisation on the elastic map presented in the internal coordinates, b)-d) Elastic map embedded in the 3-dimensional principal component space under various angles of view. Data points are in black.

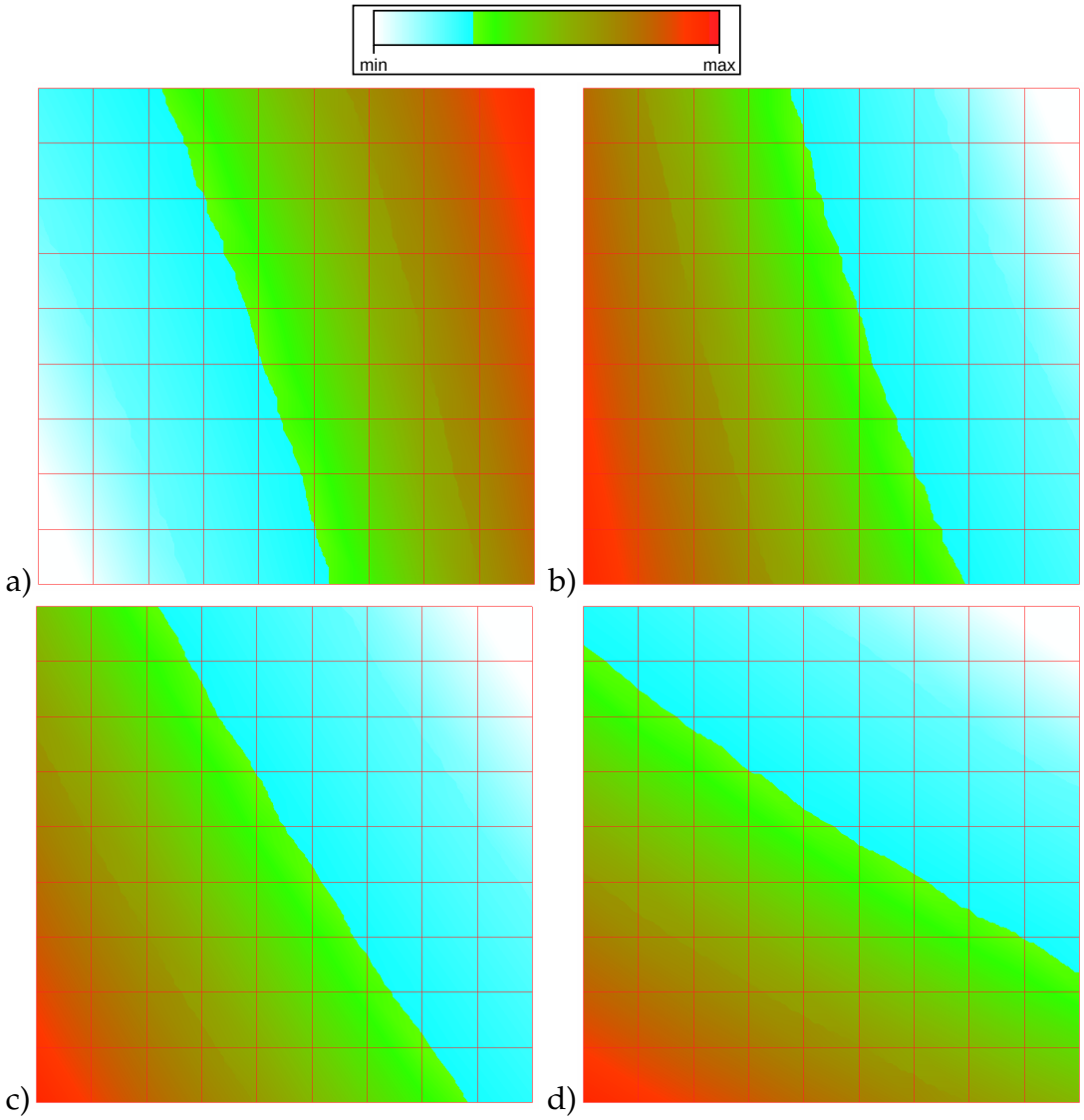

Figure 4.30. Visualisation of various functions on the elastic map (internal coordinates): a) Age, b) Imp, c) SS, d) O (age and attributes apparently correlated with age on the maps).

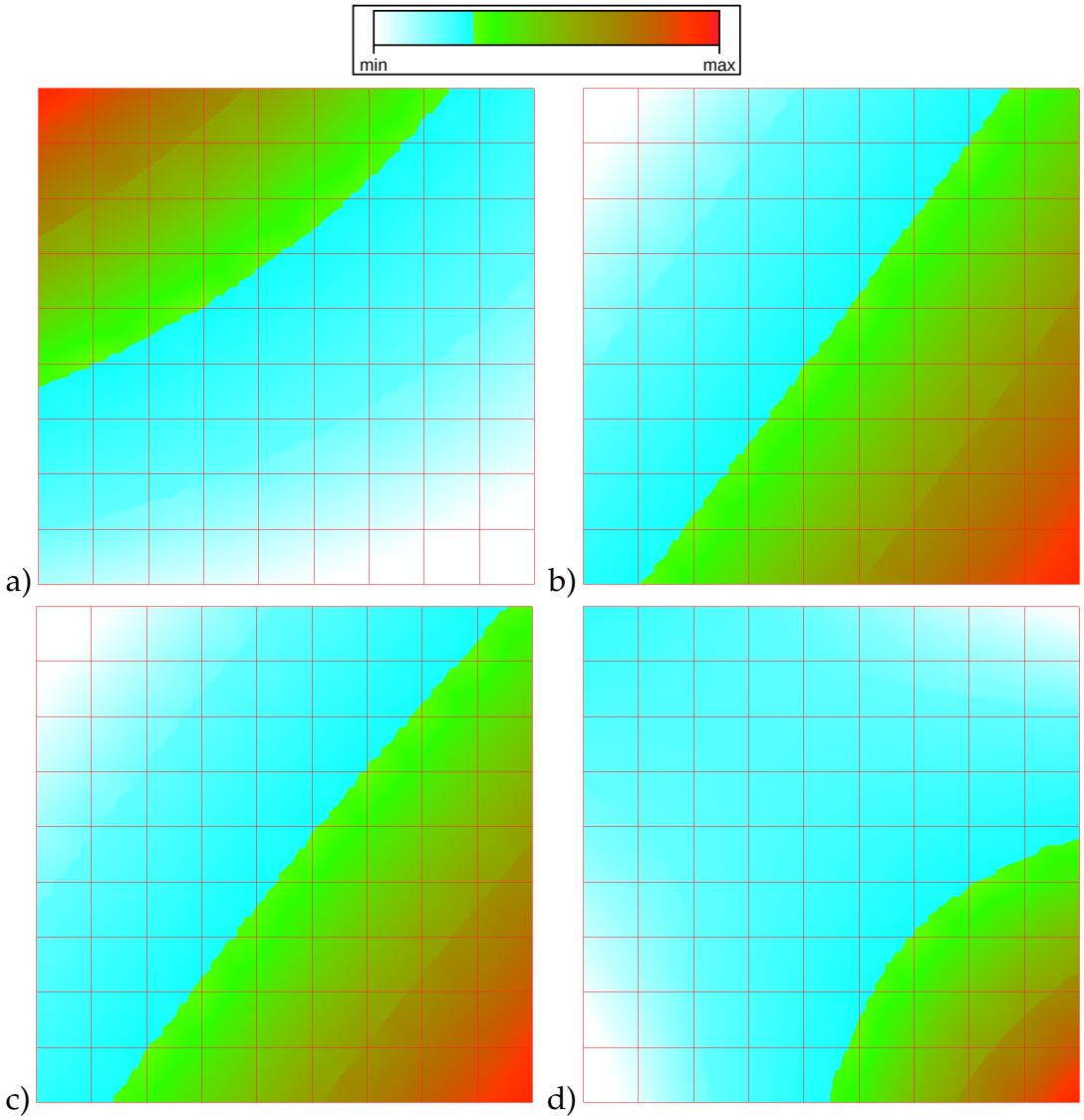

Figure 4.31. Visualisation of various functions on the elastic map (internal coordinates): a) N, b) A, c) C, d) Edu

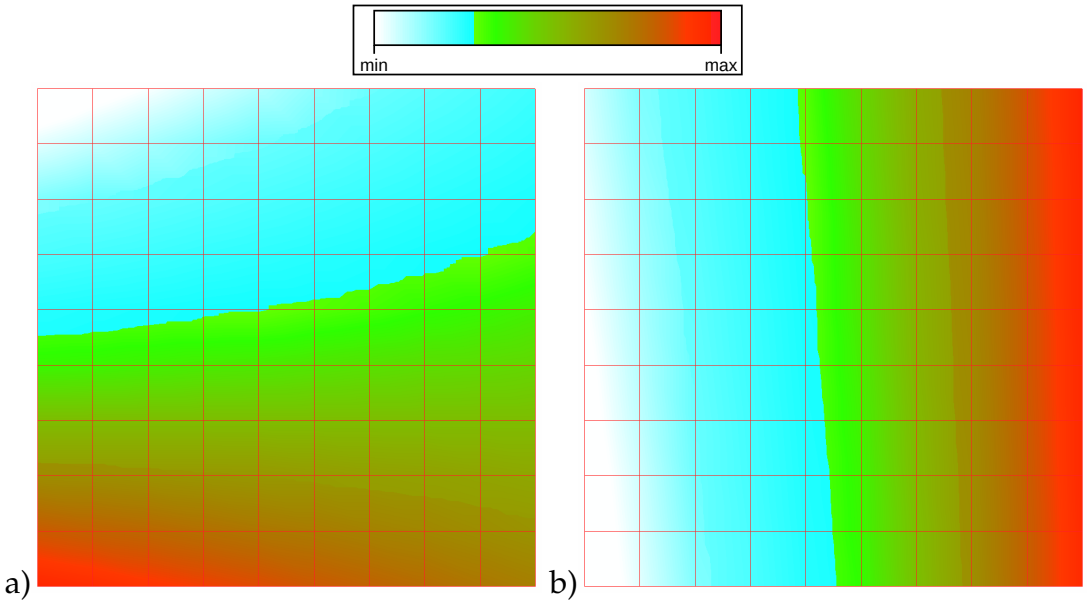

Figure 4.32. Visualisation of various functions on the elastic map (internal coordinates): a) E, b) Gndr.

the relative risk for comparison of different values of attributes.

In Fig. 4.34 we demonstrated the simplest risk maps produced by the bi-Gaussian approximation of the probability density: the densities of users and of non-users of a specific drug were approximated by two-dimensional Gaussian distributions, and then the conditional probability density ρ_u of drug use has been evaluated:

$$\rho_u = \frac{n_u N(\mu_u, S_u)}{n_u N(\mu_u, S_u) + n_{n-u} N(\mu_{n-u}, S_{n-u})},$$

where n_u and n_{n-u} are the number of users and non-users of the drug respectively, S_u and S_{n-u} are the empiric covariance matrices of the users and non-users respectively, and $N(\mu, \Sigma)$ is the normal distribution with mean μ and covariance matrix Σ. For the user/non-user separation this approximation corresponds to so-called quadratic discriminant analysis [149].

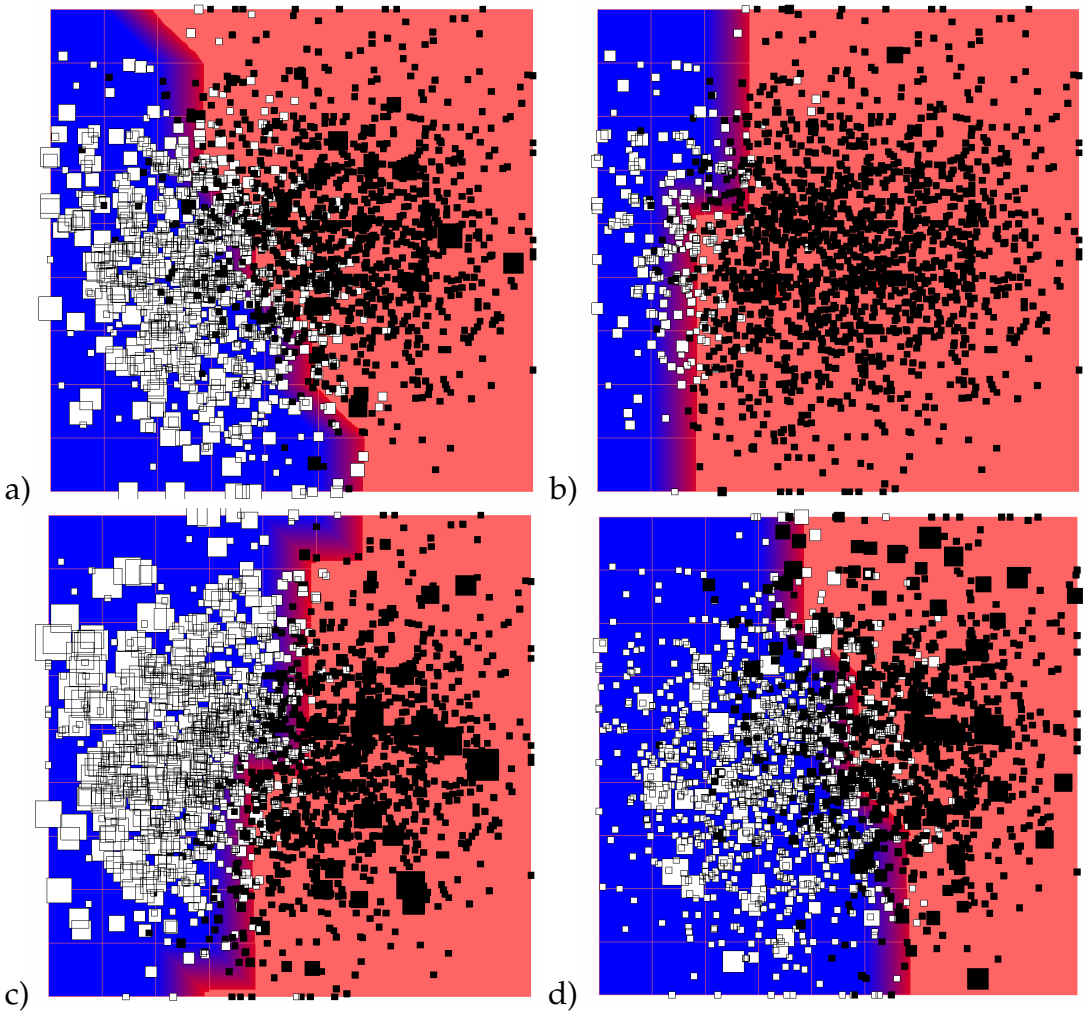

Figure 4.33. Visualisation of linear discriminant classifiers on the non-linear PCA elastic map screen: a) Ecstasy, b) Heroin, c) Benzodiazepines d) The group of 'Illicit drugs'. White squares – users, black squares – non-users. Light (red) background – LDA predicts users, dark (blue) background – LDA predicts non-users.

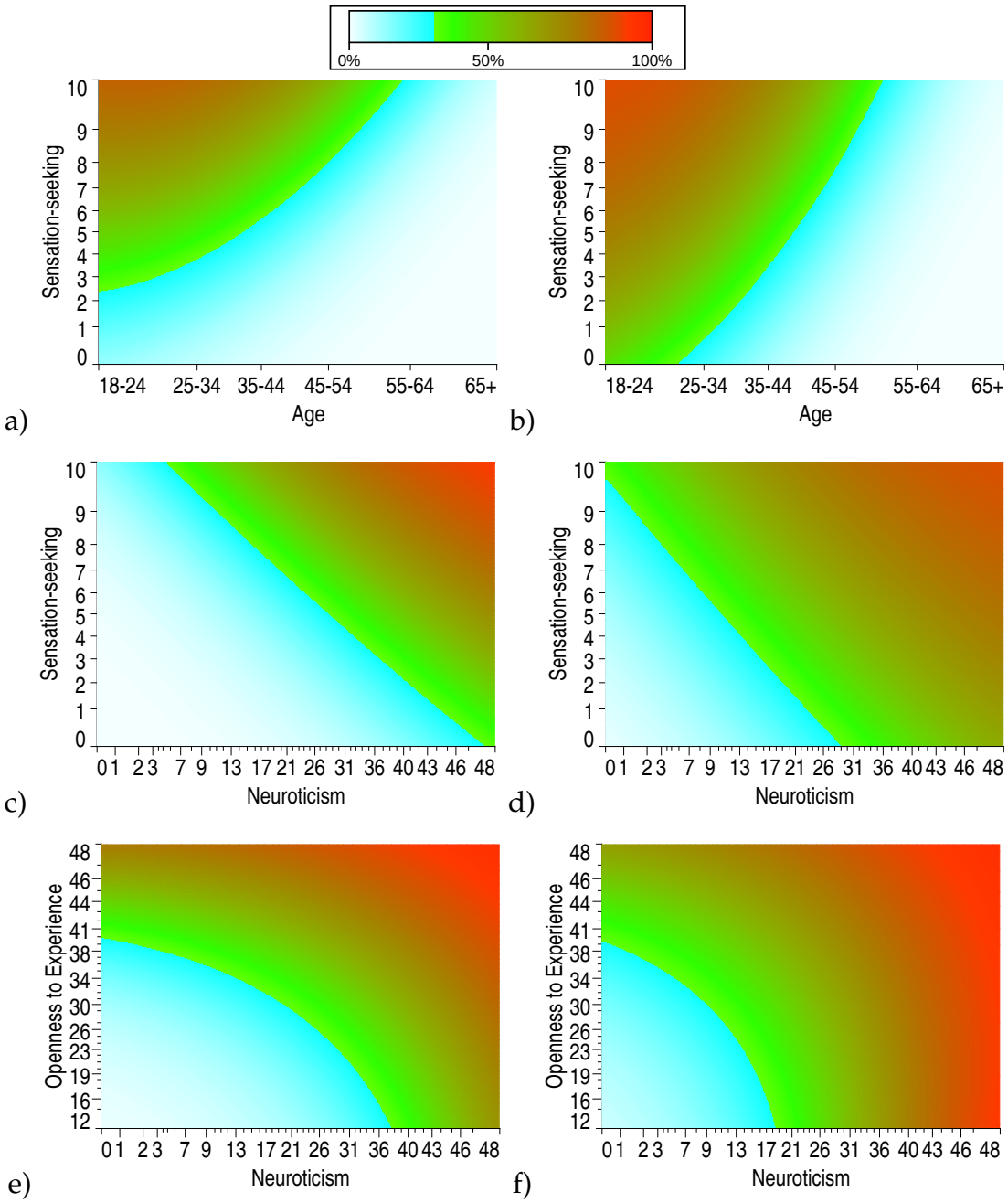

Figure 4.34. Simplest examples of risk map of: ecstasy consumption for female (a) and male (b), heroin consumption for female (c) and male (d), and benzodiazepines consumption for female (e) and male (f).

4.17 Discussion

We asked whether a psychological predisposition to drug consumption exists. Now, we can formulate the answer in brief:

- There is significant difference in psychological profiles between drug users and non-users.
- The psychological predisposition to using different drugs may be different.
- We describe the groups of drugs with correlated use (correlation pleiades) and we can lump users of these groups of drugs together for the purpose of analysis.

This study demonstrates strong correlations between personality profiles and the risk of drug use. This result partially supports observations from the previous research [10,22–24,26–29]. For example, individuals involved in the use of 'heavy' drugs like heroin and methadone are more likely to have higher scores for N, and low scores for A and C . In addition, they have significantly higher O and so the typical profile is N⇑, O⇑, A⇓, and C⇓. The profile is different for recent users (within the last year) of 'party drugs' like ecstasy, LSD, and amyl nitrite. For them, N is not high, and the typical profile is O⇑, A⇓, and C⇓.

We have analysed, in full detail, the average differences in the groups of drug users and non-users for 18 drugs (Tables 4.6, 4.7, 4.8, and 4.9). In addition to this analysis, we have achieved a much more detailed understanding of the relationship between personality traits, biographic data, and the use of individual drugs or drug clusters by an individual subject.

The database we analysed contains 1885 participants and 12 features (input attributes). These features included five personality traits (NEO-FFI-R); impulsivity (BIS-11), sensation seeking (ImpSS), level of education, age, Gndr, country of residence, and ethnicity. The data set includes information on the consumption of 18 central nervous system psychoactive drugs: alcohol, amphetamines,

amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine, as well as VSA (output attributes). This study was limited since the sample collected was biased with respect to the general population, but it remains useful for the evaluation of the risk of a person being a user of drugs.

In this analysis we used three different techniques for ranking features. After input feature ranking we excluded ethnicity and country of residence since the dataset has not enough data for most of ethnicities and countries to prove the value of this information. However, it was impossible to completely exclude the possibility that these attributes may be important risk factors. As a result, 10 input features remained: age, Edu, N, E, O, A, C, Imp, SS, and Gndr. The aim of this work was to predict the risk of drug consumption for an individual.

All input features are ordinal or nominal. To apply data mining methods which were developed for continuous input features we apply the CatPCA technique to quantify the data.

We used four different definitions of drug users based on the recency of the last consumption of drug: the decade-based, year-based, month-based and week-based user/non-user separation (Fig. 2.1). The day-based classification problem is also possible but there is not enough data on drug within the last day for most drugs.

This work have allowed us to draw a number of important conclusions about the associations between personality traits and drug use. All five personality factors are relevant traits to be taken into account when assessing the risk of an individual consuming drugs.

The mean scores for the groups of users of all 18 drugs are moderately high $(+)$ or neutral (0) for N and O, and moderately low $(-)$ for A and C. The only exception is for crack usage for the week-based classification problem, which has a moderately low $(-)$ O score (see Table 4.9 and Fig 4.9). Users of legal drugs (alcohol, chocolate, caffeine, and nicotine) have neutral A and C scores (0), other nicotine

users whose C score is moderately low $(-)$. For LSD users in the year-based classification problem and for LSD and magic mushrooms users in the week-based classification problem the A score is neutral (0).

The impact of the E score is drug specific. For example, for the decade-based user/non-user definition the E score is negatively correlated with consumption of crack, heroin, VSA, and methadone (E score is $(-)$ for their users). It is has no predictive value for other drugs for the decade-based classification (the E score for users is (0)), whereas in the year-, month-, and week-based classification problems all three possible values of E score are observed (see Tables 4.6, 4.7, 4.8 and 4.9).

We confirm the findings of previous researchers that the higher scores for N and O the lower scores for C and A lead to increased risk of drug use [30]. The O score is marked by curiosity and open-mindedness (and correlated with intelligence), and it is therefore understandable why higher O may be sometimes associated with drug use [170]. Flory et al [24] found marijuana use to be associated with lower A and C, and higher O. These findings have been partially confirmed by this study. The results improve understanding of the pathways leading to drug consumption.

It is known that significant predictors of alcohol, tobacco and marijuana use may vary according to the drug in question [171]. This study demonstrated that different attributes are important for different drugs. The detailed profiles for users and non-users of all drugs and groups of drugs are collected in Tables C.1-C.4 Using these tables we can compare, discuss, and verify almost all results and hypotheses concerning the psychological profile of drug users. For example:

1. In the paper [26] significant differences in the NEO-PI-R mean profiles of current cocaine/heroin users and non-users was found: for users, N$\Uparrow$ and C$\Downarrow$. In our table, for the month-based user definition (Table C.3), the mean profiles for cocaine users differ from non-users but the differences in A and O are also significant ($p \leq 0.001$): N$\Uparrow$, O$\Uparrow$, A$\Downarrow$, and C$\Downarrow$. For heroin our

table gives the result N⇑, E⇓, A⇓, and C⇓. The mean heroin user profile differs significantly from the profile for cocaine users: Heroine users are significantly less extravert than non-users, whereas the E-score for cocaine users is found to be slightly higher than for non-users ($p = 0.129$). Joining heroin users and cocain users in one category concealed the deviations in E-score. The mean O-score of heroin users is higher than for non-users but not significantly higher (the empirical value of the O-score for heroin users is approximately the same as for cocaine users but the significance is different because of the different sample sizes).

2. Again in [26] and in [179] significant deviations of NEO-PI-R mean profiles for current cannabis users from the means for non-users are defined as O⇑, A⇓, and C⇓. We additionally see in Table C.3 that we have N⇑. The differences in the empirical value of N-scores between cannabis users and non-users are approximately the same in our tables as in [26] (≈ 1), but the sample sizes are different and therefore, the statistical significance differs.

3. In [24] a strong difference in N between cannabis users and non-users is proposed, and no significant differences in E and O. This result strongly contradicts our observations, as well those of other earlier works [26, 137], which report a high difference in O and no (or more modest) difference in N.

4. Combinations of high and low scores in three NEO-FFI personality dimensions, neuroticism, extraversion, and conscientiousness, result in eight different personality types. Smoking, consumption of alcohol and drugs, and risky sexual behaviour were studied in a sample of 683 university students in [23]. Two types, E⇑, C⇓ (Impulsives, Hedonists) and N⇑, C⇓ (Insecures) were particularly inclined to engage in multiple, risky health behaviours, whereas the type with E⇓, C⇑ (Sceptics, Brooders) abstained from risky behaviour. Let us look at the data with the month-based user definition (Table C.3).

- Users of amphetamines, benzodiazepines, cannabis, cocaine, crack, heroine, legal highs, and nicotine belong to the type N⇑, C⇓ (Insecures) and do not belong to the type E⇑, C⇓ (Impulsives, Hedonists).
- Users of ecstasy and LSD belong to the type E⇑, C⇓ and do not belong to the type N⇑ , C⇓.
- Users of methadone belong to both types (intersection).
- It is worth mentioning that users of VSA do not belong to these types but have significant deviations in O⇑ and in A⇓ (insufficient significance of deviation in N and C may be caused by the small sample size with VSA).
- Users of magic mushrooms have significant deviations in O⇑ and in C⇓ but do not belong to both types (insignificance of deviation of E⇑ may be caused by the small number of magic mushroom users).
- Users of ketamine also have the profile O⇑ and in C⇓ (insignificance of deviation of N⇑ may be caused by the sample size).
- Users of Amyl nitrite have profile A⇓, C⇓.
- The profile with C⇑ does not exist among user profiles.

The hypothesis we make above about types of risky behaviour is partially supported by the data. Moreover, we can suggest that the type E⇑, C⇓ (Impulsives, Hedonists) is more typical among ecstasy and LSD consumers, whereas the type N⇑, A⇓ is more expected among heroin users. Detailed comparison of ecstasy and heroin users deminstrates that they are significantly different. Heroin users have higher N, lover E and A. We can also suggest that a high O-score is typical for many drug users (besides users of heroin, crack, and amyl nitrite) and therefore the O score cannot be excluded from typology of risky behaviour. Moreover, very low A⇓ is typical for VSA users. This is especially interesting because low A is the sole significant predictor of violence and is central to the dark behaviours [180].

These comments may help in the further development of the typology of risky behaviour.

We tested eight types of classifiers for each drug for the decade-based user definition. LOOCV was used to evaluate sensitivity and specificity. In this study we select the classification method which provides the maximal value of the least of sensitivity and specificity as the best one. If there is a tie on this basis, as there is in two cases, the method with maximal sum of the sensitivity and specificity is selected as the best. There were classifiers with sensitivity and specificity greater than 70% for the decade-based user/non-user separation for all drugs except magic mushrooms, alcohol, and cocaine (Table 4.18). This accuracy was unexpectedly high for this type of problem. The poorest result was obtained for the prediction of alcohol consumption.

The best set of input features was defined for each drug (Table 4.18). An exhaustive search was performed to select the most effective subset of input features, and the best data mining methods to classify users and non-users for each drug. There were 10 input features. Each of them is an important factor for risk evaluation for the use of some drugs. However, there was no single most effective classifier using all input features. The maximal number of attributes used in the best classifiers is six (out of 10) and the minimal number is two.

Table 4.18 shows the best sets of attributes for user/nonuser classification for different drugs and for the decade-based classification problem. This table together with its analogues for pleiades of drugs and all decade-year-month-week classification problems (Table 4.27) are important outputs of the analysis.

The decision tree (DT) for crack consumption used only two features, E and C, and provided sensitivity of 80.63%, and specificity of 78.57%. The DT for VSA consumption used age, Edu, E, A, C, and SS, and provided sensitivity 83.48% and specificity 77.64% (Table 4.18).

Feature age was employed in the best classifiers for 14 drugs for the decade-based classification problem, and so was a very widely used feature. Gndr was used in

the best methods for 10 drugs. We found some unexpected outcomes. For example, the fraction of females which are alcohol users is greater than the fraction of males but a majority of males consume caffeine (coffee).

Most of the features which are not used in the best classifiers are redundant but not uninformative. For example, the best classifier for ecstasy consumption used age, SS, and Gndr and had sensitivity 76.17% and specificity 77.16%. There is another DT which utilizes age, Edu, O, C, and SS with sensitivity 77.23% and specificity 75.22%, a DT with inputs age, Edu, E, O, and A, with sensitivity 73.24% and specificity 78.22%, and an advanced *k*NN classifier with inputs age, Edu, N, E, O, C, Imp, SS, and Gndr, with sensitivity 75.63% and specificity 75.75%. This means that for evaluating the risk of ecstasy usage all input attributes are informative but the required information can be extracted from a subset of attributes.

We have demonstrated that there are three groups of drugs with strongly correlated consumption. That is, drug usage has a 'modular structure'. The idea of merging correlated attributes into 'modules' is popular in biology. These modules are called the 'correlation pleiades' [43–45] (see Section 'Pleiades of drugs'). The modular structure contains three modules: the heroin pleiad, ecstasy pleiad, and benzodiazepines pleiad:

- The *Heroin pleiad (heroinPl)* includes crack, cocaine, methadone, and heroin.
- The *Ecstasy pleiad (ecstasyPl)* includes amphetamines, cannabis, cocaine, ketamine, LSD, magic mushrooms, legal highs, and ecstasy.
- The *Benzodiazepines pleiad (benzoPl)* contains methadone, amphetamines, and cocaine.

The modular structure is well represented in the correlation graph Fig 4.10. We define groups of users and non-users for each pleiad. In most of the databases the classes of users and non-users for most of the individual drugs are imbalanced (see Table 2.6), but merging the users of all drugs into one class 'drug users' does

not seem to be the best solution because of physiological, psychological and cultural differences in the usage of different drugs. We propose instead to use correlation pleiades for the analysis of drug usage as a solution to the class imbalance problem because for all three pleiades the classes of users and non-users are better balanced (Table 4.25) and the consumption of different drugs from the same pleiad is correlated.

We have applied the eight methods described in the 'Risk evaluation methods' Section and selected the best one for each problem for each of the pleiades. The results of the classifier selection are presented in Table 4.27 and the quality of the classification is high. The majority of the best classifiers for pleiades of drugs has a better accuracy than the classifiers for individual drug usage (see Tables 4.18 and 4.27). The best classifiers for pleiades of drugs use more input features than the best classifiers for the corresponding individual drugs. The classification problems for pleiades of drugs are more balanced. Therefore, we expect that the classifiers for pleiades are more robust than the classifiers for individual drugs.

The user/non-user classifiers can be also used for the formation of risk maps. Risk maps are useful tools for the visualisation of data and for generating hypotheses about the problem under consideration.

CHAPTER 5

Alternative Attributes Sets Approach (AASA)

5.1 Introduction

In this chapter, we developed a methodology for selection of several kinds of alternative sets of attributes and usage all these sets together. This methodology creates and utilises controllable multicollinearity. Multicollinearity is, at the same time, a useful and an undesirable property of data. It can be useful because it allows to correct mistakes in data and to evaluate missed data. It is undesirable because many statistical tasks become ill-conditional. We propose to optimise the multicollinearity by creation of the so-called alternative attribute sets. This approach we called '*Alternative Attributes Sets Approaches*' (AASA). AASA can find several different sets of relevant attributes such that each set can be used to solve original problems separately. Such set with the fitted model is called Alternative Attribute Set (AAS). AAS notion is based on the notion of minimal attribute set. Minimal attribute set $M(S)$ is a subset of attributes set S which provides required accuracy and does not contain any other minimal subsets. It is necessary to stress that minimal feature set is very rarely an optimal feature set. Minimal set can be found by ES, FFS, or BFS. For the minimal set selection either one or several classifier can be used. In this study for selection AAS we use six classifiers: LR,

kNN, LDA, PDFE, GM, and NB. Two kinds of AAS are defined. The first kind AAS for feature set S is the minimal set which does not contain any elements of S. The second kind AAS is the list of sets with two properties: (i) each set does not contain at least one element of set S and (ii) for each element of S there is at least one set which does not contain this element. Each first kind AAS is always the second kind AAS.

To select minimal set the required accuracy must be specified. There are many measures of classification accuracy [154]. In this study we consider that model (classifier with specified set of input features) provides accuracy α if sensitivity and specificity of the model are not less than α. The value of required accuracy is different for different data sets. A model with greater value of minimum among sensitivity and specificity is considered as a better model. If two models have the same value of minimum among sensitivity and specificity, then we select model with greater sum of the sensitivity and specificity. LOOCV [120] is used for all tests in this study.

We applied AASA for three real life datasets which have different number of records and different dimensions and are taken from different application areas. The first database is 'Drug consumption' (psychology, [2], section 'Database' of chapter 2). The second database is 'USA president elections' (politics, [4]). The third dataset is 'Breast cancer' (medicine, [61]). Required accuracies for minimal set identification are defined as 65% for drug consumption, 75% for USA president elections, and 95% for breast cancer dataset.

For each classification problem and each FS method we define five candidates to the best model. The first candidate is the best model among models which are tested by basic FS method (ES, FFS, or BFS). It is not AASA model and is used as a reference point for comparison with AASA models. We call it FS model. The other four models are AASA models: union model for the first kind AAS, ensemble model for the first kind AAS, union model for the second kind AAS, and ensemble model for the second kind AAS. Then we select the best model

among FS models and best model among AASA models. We found that the best AASA model usually is much better than the best FS model. In most cases the best AASA model is ensemble model for first or second kind AAS.

5.2 Standard feature selection approaches

Let us consider problem of classification with input feature set $F=\{f_1,\ldots,f_m\}$ where m is a number of input features. Let $X=\{x_1,x_2,\ldots,x_n\}$ is data matrix where n is the number of observations. Each record x_i is a vector $x_i=(x_{i1},x_{i2},\ldots,x_{im})$, where x_{ij} is the value of attribute f_j in observation x_i. The problem of selection of the best subset of F is well-known [54–56, 172].

In this work we consider the wrapper FS [51, 55, 57] approaches only because it is developed to select the features in accordance with accuracy of problem solution. The feature selection for classification problem is defined as identifying of the minimum size subset of features that provide the maximal predictive performance [54, 55]. There are three widely used methods of embedded FS:

- ES or brute-force search is the method exhaustively evaluates and tests all possible subsets of the input features, and then selects the best subset [54–56]. Clearly, the ES's computational cost is prohibitively high.
- FFS or greedy forward selection is greedy method which initially searches for the best feature f_1^*, then search the best pair of features (f_1^*,f_2^*) with f_1^* which is found on the first step, then search the best triple (f_1^*,f_2^*,f_3^*) with previously found f_1^* and f_2^*, etc. [54–56].
- BFS or greedy backward elimination is the greedy method which initially search the best subsets $F\setminus\{f_1^*\}$, then search the best subset $F\setminus\{f_1^*,f_2^*\}$ with f_1^* which is found on the first step, then search the best subset $F\setminus\{f_1^*,f_2^*,f_3^*\}$ with previously found f_1^* and f_2^*, and so on [54–56].

Table 5.1. Comparison time costs of three feature selection methods

m	Exhaustive	Forward	Backward
5	240	105	155
10	28,160	1,210	2,410
15	1,966,080	5,440	12,265
20	110,100,480	16,170	38,970
50	7.2×10^{17}	563,550	1,543,550
100	3.2×10^{33}	8,670,850	24,840,850

5.3 Time consumption of three classical algorithms

Three classical algorithms have different time costs. Let us consider the logistic regression as used classifier. Time for fitting one logistic regression model with k input features is proportional to k^2. It means that for ES time cost is proportional to

$$t_{ES} = \sum_{k=1}^{m} \binom{m}{k} k^2 = m(m+1)2^{m-2}$$

For forward feature selection time costs is proportional to

$$t_{FS} = \sum_{k=1}^{m} (m-k+1)k^2 = \frac{m(m+1)^2(m+2)}{12}$$

For backward feature selection it is necessary to spend time which is proportional to

$$t_{BS} = m^2 + \sum_{k=1}^{m-1} (m-k+1)(m-k)^2 = \frac{m}{12}(3m^3 - 2m^2 + 9m + 2)$$

Comparison of time costs of three feature selection methods is presented in the Table 5.1. Table 5.1 shows that ES can be used on data sets with a small number of features ($m \leq 20$). Unfortunately FFS and BFS can find quasi optimal feature set only. It is argued that FFS is computationally more efficient than BFS to search nested subsets of attributes. On the other hand, the protectors of BFS argue that weaker subsets are found by FFS because the importance of features is not measured in the context of other features which are not included yet.

5.4 Alternative Attributes Set (AAS)

Usage of all accessible input features is not the best practice [55, 56] because of multicollinearity problem [154, 173]. From the other side the models with minimal set of attributes are non-robust with respect to any errors in input data. To resolve this trade-off problem many FS was developed [56,57]. The purpose of FS is to remove irrelevant and redundant features [54,56]. In accordance with [55,56] feature is irrelevant if it does not provide useful information and can be the sources of noise. Exclusion of irrelevant features is unconditional. The different case is redundant features. The main difference is that feature can be redundant with respect to specified set of attributes. It means that this attribute cannot add new information in comparison with information which is provided by other attributes. The same attribute can be non-redundant with respect to another set of attributes. For example, for heroin consumption problem feature Gndr is redundant with respect to set of two attributes: E and SS but it is not redundant with respect to set of two other features: N and Imp (see Table D.1).

The elementary operation of AASA is to search of minimal set of attributes which can be found by ES, FFS, and BFS. Required accuracy also has to be specified before it. The main idea of AASA is to select minimal sets of attributes which can separately solve problem with required accuracy. It means that elements of one set are redundant for another set. Each such set is AAS.

Definition 1. Minimal feature set $M(S)$ is a subset of attributes set S which provide required accuracy and does not contain any other minimal subsets.

For example, the minimal feature set for FFS is the first subset which provides required accuracy. While, for BFS it is the last subset which provides required accuracy. Minimal set is not unique. For example, there are 47 minimal sets for heroin consumption problem (see Table D.1).

We define two types of AAS: AAS of the first and second kinds.

Definition 2. The first kind AAS for feature set S is the minimal set which does

not contain any elements of S $A_1(S)$=$M(F \setminus S)$, where $A \setminus B = \{x \in A, x \notin B\}$ is relative complement of B in A or set-theoretic difference of A and B.

Definition 3. The second kind AAS for feature set S is the list of sets with two properties: (i) each set does not contain at least one element of set S and (ii) for each element of S there is at least one set which does not contain this element.

The main difference of this two kinds of AAS is in object for substitution. The first kind AAS is substituted all information which is provided by all features of set S. The second kind AAS is substituted the information which is provided by one or several features from S. Each first kind AAS is always the second kind AAS. Difference between two kinds of AAS causes existence of several AAS of the first kind for the same set S. The second AAS can be searched as $A_1(S \bigcup A_1(S))$, the third AAS of the first kind can be searched as $A_1(S \bigcup A_1(S) \bigcup A_1(S \bigcup A_1(S))$ etc. The AAS of second kind usually is a list of several sets and we cannot search 'the second AAS of the second kind' by definition.

We used the several classification methods to finding AASA model. We can use several classifiers for two purposes: (i) for AAS finding (ii) for forming of the final model. Formed union model is always one classification model and usage of several classifier can only help to select the best one. Ensemble model created with usage of several classifiers can include different classifiers in the first layer (see Figure 5.3).

AASA is developed to create more robust and/or more accurate model. There are two ways to use all AAS for model construction. If there are several sets which are alternative to each other (several AAS), then we can unite feature sets of all AAS in one set and construct a model with this set of features. Let us call such model *'union model'*. Procedure of union model forming is depicted in Figure 5.1. Outcome of this procedure is usual classifier which uses part of attributes only. It means that for union models AASA is another FS method. Another way to utilize several AAS is a creation of two layer model: first layer contains classifiers for each AAS separately and second layer contains a classifier which uses

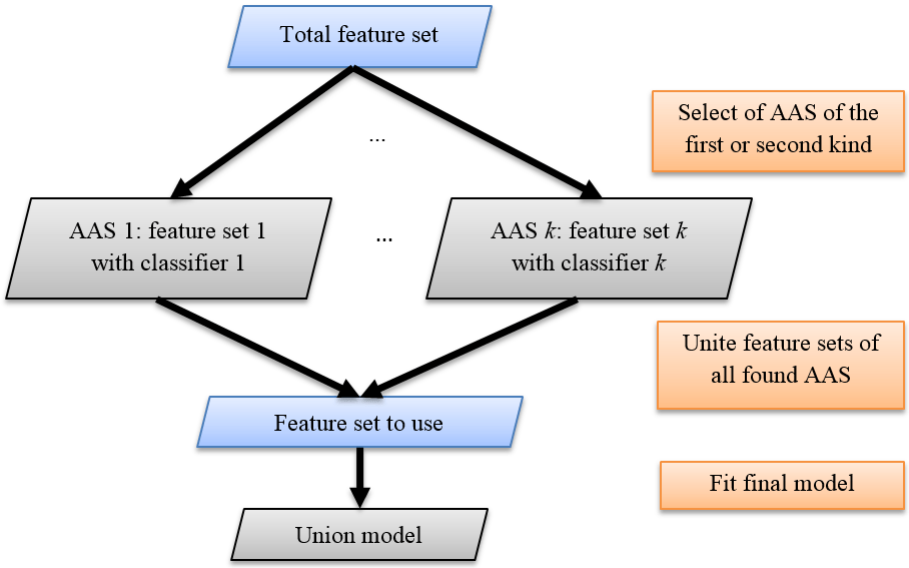

Figure 5.1. Construction of union model.

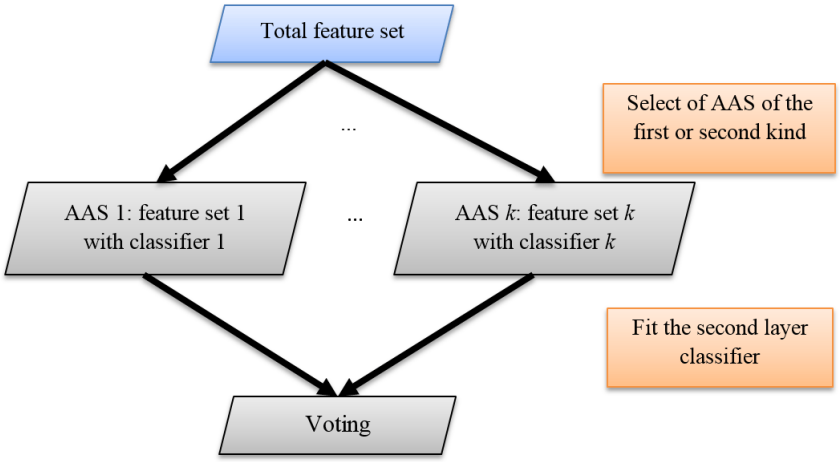

Figure 5.2. Construction of ensemble model.

outcomes of the first layer classifiers as input features. We call such model *'ensemble model'* [59, 60]. Procedure of the ensemble model creation is depicted in Figure 5.2. We consider linear ensembles only. Linear ensemble outcome is defined as $\sum w_i e_i \; / \; \sum w_i$, where e_i is outcome of ith classifier of the first layer and w_i is its weight. We used several techniques to define weights of ensemble model: simple voting, Nelder-Mead, Luus-Jaakola, pattern search, and linear regression. These techniques are demonstrated in chapter 3 for k AAS.

Structure of the ensemble model is presented in Figure 5.3

We identified AASA by flowchart based on FS methods (ES, FFS, BFS) by using

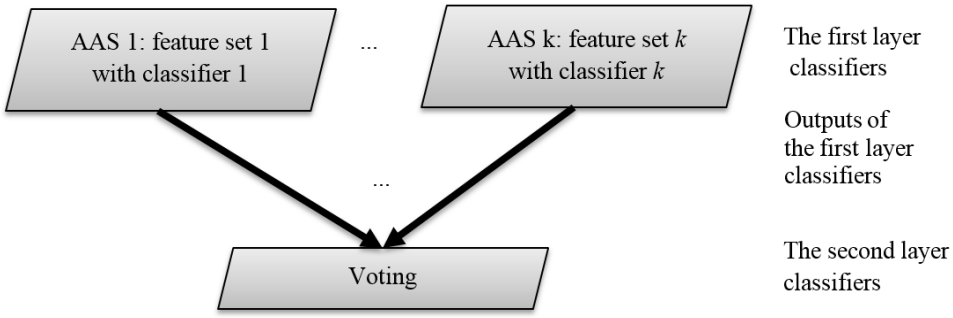

Figure 5.3. Structure of ensemble model.

the classifiers. All flowcharts of AASA for feature selection are presented in Appendix B.

5.5 AASA with usage of several classifiers

It is well known that sometimes classifier combined from classifiers of several types (ensemble model) has accuracy better than each of original classifiers [60, 108]. We can use several classifiers for two purposes: (i) for AAS finding and (ii) for forming of the final model. Formed union model always is one classification model and usage of several classifier can only help to select the best one. Ensemble model created with usage of several classifiers can include different classifiers in the first layer (see Figure 5.3). Definition 1 of the minimal set for the case of several classifiers can be clarified as: minimal feature set $M(S)$ is a subset of attributes set S which provide required accuracy for at least one classifier and does not contain any other minimal subsets. This clarification is useful but is not necessary. Definitions of the first and second kind AAS are exactly the same as for the usage of one classifier.

5.6 AASA for drug consumption dataset

To apply AASA for drug consumption dataset we select three drugs: heroin, ecstasy, and cannabis. Also we select the decade-based user/non-user separation.

In chapter 4 we show that the accuracy of DT models for these drugs is: 79% for cannabis, 76% for ecstasy, and 73% for heroin which ES was performed to select the most effective subset of input features (see Section 'Selection of the best classifiers for the decade-based classification problem' and Table 4.18). Recall that accuracy of model in this study is minimum among sensitivity and specificity. To find any AAS it is necessary to select less accuracy. For drug consumption dataset we select required accuracy equals 65%. We illustrate AASA for this three drugs to find the most accurate model by usage LR as the only classifier and all three FS techniques: ES, FFS, and BFS. Then we examine this approach for ecstasy and cannabis consumption by usage of six classifiers and all three FS techniques.

5.6.1 AASA for heroin usage by using LR and ES, FFS, and BFS

This subsection contains detailed description of AAS selection by LR and each of the FS methods. The best model which is found by ES has sensitivity 73.11% and specificity 69.99% and use seven attributes: age, Edu, N, E, O, Imp, and SS.

FFS based the first kind AAS

In accordance with definition of minimal set, it is the first set selected by FFS which provides required accuracy. Table 5.2 illustrates the result of FFS by LR. In Table 5.2 the green background highlights the best LR model based on FFS and the yellow background highlights the first (minimal) feature set which provides required accuracy. This set which includes E and SS we call M_1.

To find the first kind AAS for M_1 we repeat FFS for set of all features exclude E and SS. Protocol of FFS is presented in the Table 5.3 . The minimal set M_2 is highlighted by yellow background and contains Edu, N and O. Set M_2 is the first AAS of the first kind for set M_1

To find the second AAS for M_1 we have to exclude features of M_1 and M_2 from set of used features: we repeat FFS for set of all features exclude E, SS, Edu, N, and

Table 5.2. Protocol of FFS by LR for heroin consumption, column '#' contains number of used features, 'X' means used input feature

#	Age	Edu	N	E	O	A	C	Imp	SS	Gndr	Sn (%)	Sp (%)	Sum (%)
1									X		68.40	62.46	130.86
2				X					X		69.81	67.36	137.18
3			X	X					X		69.81	66.95	136.76
4			X	X					X	X	70.76	68.08	138.84
5			X	X		X			X	X	70.76	68.56	139.31
6			X	X	X	X			X	X	70.76	68.38	139.14
7			X	X	X	X		X	X	X	71.23	69.22	140.44
8		X	X	X	X	X		X	X	X	72.64	69.04	141.68
9		X	X	X	X	X	X	X	X	X	73.11	68.98	142.09
10	X	X	X	X	X	X	X	X	X	X	69.34	68.98	138.32

Table 5.3. Protocol of FFS by LR for heroin consumption without features E and SS, column '#' contains number of used features, 'X' means used input feature

#	Age	Edu	N	O	A	C	Imp	Gndr	Sn (%)	Sp (%)	Sum (%)
1		X							62.26	62.28	124.55
2		X		X					66.04	63.60	129.64
3		X	X	X					67.45	67.36	134.82
4	X	X	X	X					70.28	68.14	138.42
5	X	X	X	X	X				68.87	69.22	138.09
6	X	X	X	X	X		X		69.34	68.62	137.96
7	X	X	X	X	X		X	X	70.28	69.28	139.56
8	X	X	X	X	X	X	X	X	70.76	69.04	139.79

Table 5.4. Protocol of FFS by LR for heroin consumption without features Edu, E, N, O, and SS, column '#' contains number of used features, 'X' means used input feature

#	Age	A	C	Imp	Gndr	Sn (%)	Sp (%)	Sum (%)
1			X			62.74	60.97	123.70
2			X	X		62.74	65.51	128.25
3	X		X	X		67.93	65.51	133.44
4	X	X	X	X		67.93	66.83	134.75
5	X	X	X	X	X	68.87	65.99	134.86

Table 5.5. Protocol of FFS by LR for heroin consumption with features A and Gndr, column '#' contains number of used features, 'X' means used input feature

#	A	Gndr	Sn (%)	Sp (%)	Sum (%)
1	X		56.60	60.79	117.39
2	X	X	67.45	61.33	128.78

O. Protocol of FFS is presented in the Table 5.4. The minimal set M_3 is highlighted by yellow background and contains age, Imp, and C. Set M_3 is the second AAS of first kind for set M_1.

To find the third AAS for M_1 we repeat FFS for feature set after exclusion of M_1, M_2, and M_3. Protocol of FFS for feature set without E, SS, Edu, N, O, Imp, age, and C is presented in Table 5.5. Both feature sets in Table 5.5 do not satisfy restriction to have at least 65% of sensitivity and specificity. As a result there are three sets which are AAS of each other: M_1, M_2, and M_3. Union of these sets contains eight attributes: age, O, Imp, SS, C, E, Edu, and N. The LR union model has sensitivity 73.11% and specificity 69.76%. The ensemble model for these three AAS has sensitivity 76.89% and specificity 69.58%.

Further protocols of FFS do not included into thesis.

FFS based the second kind AAS

Table 5.2 shows that the minimal set is M_1 and contains E, SS. We have to repeat FFS for feature set without each of elements of M_1 separately. We reapply FFS for all features without E to find minimal feature set which does not contain E. Found AAS is denoted M_2 and contains N, and SS. We repeat FFS for all features without SS and find AAS M_3 which contains Edu, N, and O. Sets M_1, M_2 and

Table 5.6. Protocol of BFS by LR for heroin consumption, column '#' contains number of used features, 'X' means used input feature

#	Age	Edu	N	E	O	A	C	Imp	SS	Gndr	Sn (%)	Sp (%)	Sum(%)
10	X	X	X	X	X	X	X	X	X	X	69.34	68.98	138.32
9	X	X	X		X	X	X	X	X	X	70.76	69.40	140.15
8	X	X	X		X	X	X	X	X		70.76	69.82	140.57
7	X	X	X		X	X		X	X		69.81	69.58	139.39
6	X	X	X		X	X			X		69.34	69.34	138.68
5		X	X		X	X			X		71.70	69.10	140.80
4		X	X			X			X		70.76	67.96	138.72
3		X	X						X		71.23	67.42	138.65
2			X						X		70.28	66.11	136.39
1									X		68.40	62.46	130.86

M_3 include five attributes E, Edu, SS, N, and O. The union model has sensitivity 68.87% and specificity 67.78%. The ensemble model for these three AAS has sensitivity 73.58% and specificity 68.38%.

BFS based the first kind AAS

The minimal set in the protocol BFS is the last set which provides required accuracy. Table 5.6 presents protocol of BFS by LR. In Table 5.6 the green background highlights the most accurate feature set and the yellow background highlights the minimal feature set. This set is called M_1 and includes N and SS.

To find the first kind AAS we repeat BFS for set of all features exclude N and SS. Protocol of BFS is presented in the Table 5.7. The minimal set M_2 is highlighted by yellow background and contains Edu, Imp, and O. Set M_2 is the first alternative of first kind of set M_1.

To find the second AAS we repeat BFS for set of all features exclude N, SS, Edu, Imp, and O. Protocol of BFS is presented in the Table 5.8. The minimal set M_3 is highlighted by yellow background and contains age, A, and C. Set M_3 is the second AAS of first kind for set M_1.

Protocol of BFS for feature set without N, SS, Edu, A, O, age, C, and Imp is presented in Table 5.9. Both feature sets in the table do not satisfy restriction to have

Table 5.7. Protocol of BFS by LR for heroin consumption without features N and SS, column '#' contains number of used features, 'X' means used input feature

#	Age	Edu	E	O	A	C	Imp	Gndr	Sn (%)	Sp (%)	Sum (%)
8	X	X	X	X	X	X	X	X	71.23	68.14	139.37
7	X	X	X	X	X	X	X		70.28	68.26	138.54
6	X	X	X	X	X	X			71.23	68.68	139.91
5	X	X		X	X	X			69.34	67.96	137.30
4		X		X	X	X			69.81	67.66	137.47
3		X		X		X			69.81	66.35	136.16
2				X		X			64.62	63.84	128.46
1				X					61.32	60.97	122.29

Table 5.8. Protocol of BFS by LR for heroin consumption without features N, SS, Edu, Imp, and O, column '#' contains number of used features, 'X' means used input feature

#	Age	E	A	C	Gndr	Sn (%)	Sp (%)	Sum (%)
5	X	X	X	X	X	67.93	65.51	133.44
4	X		X	X	X	69.34	65.45	134.79
3	X		X	X		67.45	65.33	132.79
2			X	X		62.74	62.52	125.26
1				X		62.74	60.97	123.70

at least 65% of sensitivity and specificity.

As a result we have three sets which are AAS of each other: M_1, M_2 and M_3. Union of these sets contains eight attributes: age, Edu, N, O, A, C, Imp, and SS. The union model has sensitivity 70.76% and specificity 69.82%. The ensemble model for these three AAS contain eight attributes and has sensitivity 74.53% and specificity 68.38%.

Further protocols of BFS do not included into paper.

BFS based the second kind AAS

Table 5.6 shows that the minimal set is M_1 and contains N, SS. We repeat BFS for all features without N and find the AAS M_2 which contains E, and SS. We

Table 5.9. Protocol of BFS by LR for heroin consumption with features E and Gndr, column '#' contains number of used features, 'X' means used input feature

#	E	Gndr	Sn (%)	Sp (%)	Sum (%)
1	X		68.40	52.78	121.18
2	X	X	52.83	54.09	106.93.78

repeat this procedure for all features without SS as well and find AAS M_3 which contains N, and O. Sets M_1, M_2, and M_3 include four attributes: N, SS, E, and O. The union model has sensitivity 72.17% and specificity 67.07%. The ensemble model for these three AAS has sensitivity 73.58% and specificity 67.72%.

ES based the first kind AAS

The best FS model for ES is selected by LR and has sensitivity 73.11% and specificity 69.99%. This model use seven attributes: age, Edu, N, E, O, Imp, and SS. For heroin consumption there are 47 minimal feature sets (see Table D.1 in Appendix). Since we have several minimal sets which are satisfying all requirements then we have to select the best set in accordance with the best model criterion. Table D.1 shows that the best minimal set in accordance with the criterion is M_1 and contains E and SS. There are 28 minimal sets which can be selected as the first AAS of the first kind for set M_1: M_2, M_3, M_4, M_5, M_6, M_8, M_9, M_{10}, M_{13}, M_{14}, M_{15}, M_{16}, M_{20}, M_{21}, M_{26}, M_{27}, M_{28}, M_{29}, M_{32}, M_{33}, M_{36}, M_{38}, M_{39}, M_{40}, M_{42}, M_{44}, M_{45}, M_{46}. Set M_2 has the best accuracy and contains N, Imp, and Gndr. It is the first AAS of the first kind for set M_1. There are four minimal sets which do not contain any of the attributes in sets M_1 and M_2: $M_{26}, M_{28}, M_{39}, M_{42}$. The second AAS of the first kind is M_{26} which contains age, A, and C. There is no AAS for sets M_1, M_2,and M_{26}.

As a result there are three sets which are AAS of each other: M_1, M_2 and M_{26}. Union of these sets contains eight input features: age, N, E, A, C, SS, Imp, and Gndr. This union model has sensitivity 70.28% and specificity 68.80%. The ensemble model for these three AAS has sensitivity 71.23% and specificity 72.15%.

ES based the second kind AAS

Table D.1 shows that there are 47 minimal sets. Minimal set with the best accuracy of LR model is M_1. All other 46 sets are the second kind AAS for M_1 in

accordance with definition. Union of these sets contains all features. The union model has sensitivity 69.34% and specificity 68.98%. Since we can use all 47 sets as second kind AAS, we used outputs of LR models for all sets as input feature to form ensemble model. Then we apply the FS methods BFS and FFS to select part of AAS. The ES is too expansive for 47 features. The best model is found by secondary BFS. This model contains all features and has sensitivity 71.55% and specificity 71.70%.

Comparison of optimal models for ES, FFS, and BFS and all AASA models on base of ES, FFS and BFS

The best LR models for ES, FFS, and BFS and all AASA results for heroin consumption are presented in Table 5.10. The fundamental brick of AASA is minimal model. To illustrate the relation between minimal model and AAS models, we include minimal models for each FS method to Table 5.10. Table 5.10 shows that all three the minimal feature sets contain two features. The minimal sets for ES and FFS are the same and contain E and SS. The minimal set for BFS contains N and SS. The minimal model based on ES and FFS is slightly better than the BFS minimal model. All minimal models are worse than the best model for each FS method.

The best AASA model is ensemble model of the first kind for ES and has sensitivity 71.23% and specificity 72.15%. The second best AASA model is ensemble model of the second kind for ES and has sensitivity 71.55% and specificity 71.70%. These two models have accuracies which are essentially better than the best ES model.

Graphs of accuracy versus number of used features for different FS methods are presented in Figure 5.4 (for ES the best accuracy is depicted). We can see that all AASA results have approximately the same or better accuracy compared with models with the same number of features. Figure 5.5 shows the sets of features which are used in the best ES model, in the best ensemble and union models for

Table 5.10. The results of LR models selected by ES, FFS, and BFS and AASA for heroin consumption, column '#' contains number of used features

Model	#	Sn (%)	Sp (%)	Sum(%)
Best ES	7	73.11	69.99	143.11
Best FFS	7	71.23	69.22	140.44
Best BFS	8	70.76	69.82	140.57
ES Minimal	2	69.81	67.36	137.18
FFS Minimal	2	69.81	67.36	137.18
BFS Minimal	2	70.28	66.11	136.39
ES Kind1. Union	8	70.28	68.80	139.08
FFS Kind1. Union	8	73.11	69.76	142.87
BFS Kind1. Union	8	70.76	69.82	140.57
ES Kind 1. Ensemble	8	71.23	72.15	143.37
FFS Kind 1. Ensemble	8	76.89	69.58	146.46
BFS Kind 1. Ensemble	8	74.53	68.38	142.91
ES Kind 2. Union	10	69.34	68.98	138.32
FFS Kind 2. Union	5	68.87	67.78	136.65
BFS Kind 2. Union	4	72.17	67.07	139.24
ES Kind 2. Ensemble	10	71.55	71.70	143.25
FFS Kind 2. Ensemble	5	73.58	68.38	141.97
BFS Kind 2. Ensemble	4	73.58	67.72	141.31

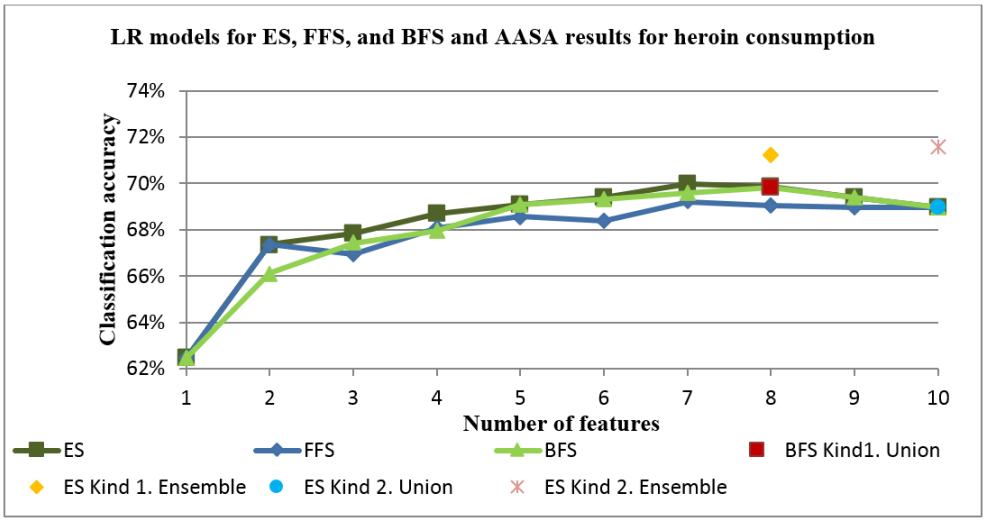

Figure 5.4. Comparisons of LR models selected by ES, FFS, and BFS and AAS results for heroin consumption. ES kind 1 Ensemble and ES kind 2 Ensemble model are the best models.

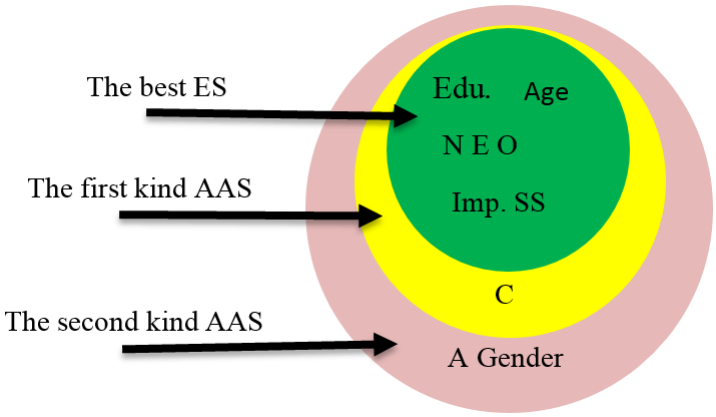

Figure 5.5. Venn diagram for features which are used for the best ES model, the first kind ensemble and union models and the second kind ensemble and union models for heroin consumption.

the first kind of AAS and for the second kind of AAS.

5.6.2 AASA for cannabis and ecstasy consumption by usage LR and ES, FFS, and BFS

For ecstasy consumption the best FS LR model is found for ES and BFS. This model contains seven features and has sensitivity 74.83% and specificity 74.52%. There are 23 minimal feature sets with appropriate accuracy. These sets are presented in Table D.3 in Appendix. The best LR models for ES, FFS and BFS and results of AASA are presented in Table 5.11.

For ecstasy consumption the first and second kind ensemble models have slightly better accuracy than the best FS model. It is necessary to stress that ensemble models are usually more robust [60, 108]. Numbers of used features are different: the best FS model includes seven attributes, the first kind ensemble model includes four attributes, and the second kind ensemble model includes nine attribute.

Graphs of accuracy versus number of used features for ES, FFS, and BFS and AASA results are presented in Figure 5.6. Figure 5.6 shows that accuracy of AASA models at least not worse than accuracy of other models with the same

Table 5.11. The results of LR models selected by ES, FFS, and BFS and AASA for ecstasy consumption, column '#' contains number of used features

Model	#	Sn (%)	Sp (%)	Sum(%)
Best ES	7	74.83	74.52	149.35
Best FFS	7	74.83	74.34	149.17
Best BFS	7	74.83	74.52	149.35
ES Minimal	2	74.70	70.64	145.34
FFS Minimal	2	71.24	69.14	140.37
BFS Minimal	2	74.70	70.64	145.34
ES Kind 1. Union	4	73.37	72.58	145.94
FFS Kind 1. Union	4	73.50	74.07	147.58
BFS Kind 1. Union	4	73.50	73.90	147.40
ES Kind 1. Ensemble	4	74.83	73.37	148.20
FFS Kind 1. Ensemble	4	74.43	74.43	148.86
BFS Kind 1. Ensemble	4	74.97	74.78	149.75
ES Kind 2. Union	10	74.57	74.16	148.73
FFS Kind 2. Union	3	74.17	71.78	145.95
BFS Kind 2. Union	4	73.37	72.58	145.94
ES Kind 2. Ensemble	9	75.10	75.13	150.23
FFS Kind 2. Ensemble	3	74.43	73.54	147.98
BFS Kind 2. Ensemble	4	74.97	72.49	147.45

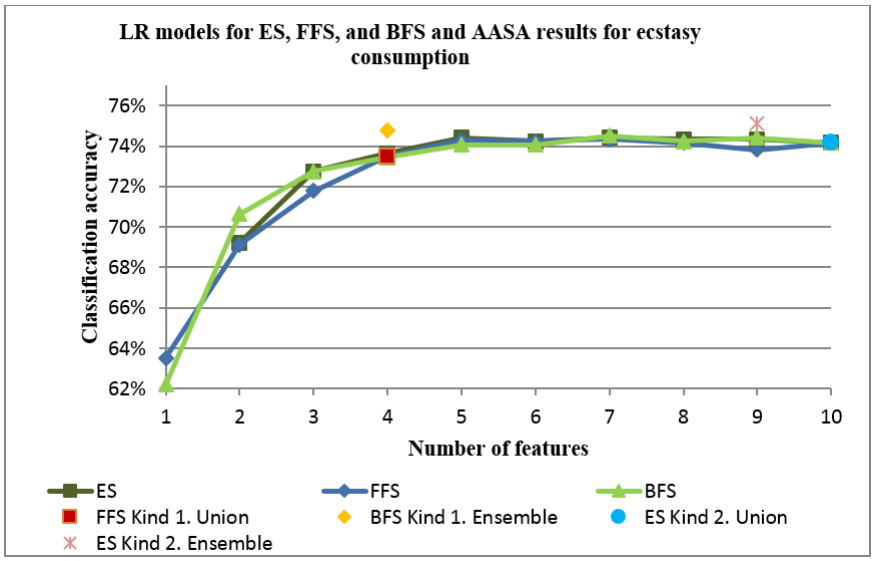

Figure 5.6. Comparisons of LR models selected by ES, FFS, and BFS and AAS results for ecstasy consumption. BFS Kind 1 Ensemble and ES Kind 2 Ensemble are the best model.

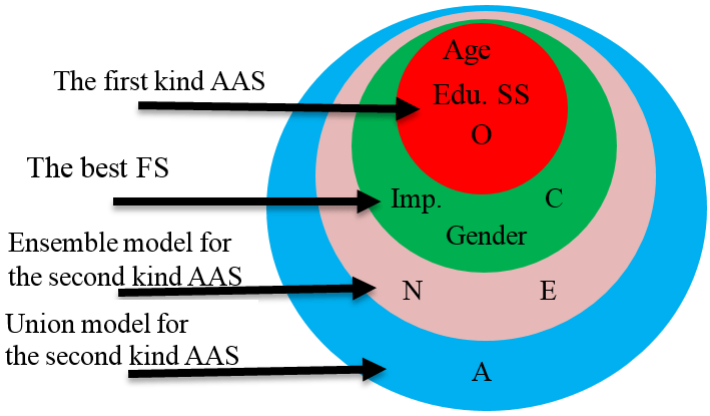

Figure 5.7. Venn diagram for features which are used for the best ES model, the first kind ensemble and union models and the second kind ensemble and union models for ecstasy consumption.

number of attributes. Figure 5.7 shows the venn diagram to sets of features which are used in the best FS model, in the best ensemble and union models for the first and second kind of AAS.

For cannabis consumption the best FS LR model is found by ES and BFS. This model has sensitivity 76.68% and specificity 77.42% and uses eight attributes. There are 9 minimal feature sets with accuracy which is not worse than required. These sets are presented in Table D.2 in Appendix. The best LR models for ES, FFS and BFS and results of AASA are presented in Table 5.12. The best FS model is better than all AASA models. Let us consider the best ensemble model (yellow background in Table 5.12) and the best union model (red background in Table 5.12) for the first kind of AAS and the best ensemble model (pink background in Table 5.12) and the best union model (blue background in Table 5.12) for the second kind of AAS. These four AASA models have greater specificity and sum of sensitivity and specificity in comparison with the best FS model. Three of these four models also have less number of used attributes.

Table 5.12. The results of LR models selected by ES, FFS, and BFS and AASA for cannabis consumption, column '#' contains number of used features

Model	#	Sn (%)	Sp (%)	Sum(%)
Best ES	8	76.68	77.42	154.10
Best FFS	9	75.89	79.84	155.73
Best BFS	8	76.68	77.42	154.10
ES Minimal	1	73.12	67.90	141.03
FFS Minimal	1	73.12	67.90	141.03
BFS Minimal	1	73.12	67.90	141.03
ES Kind 1. Union	6	73.99	79.68	153.67
FFS Kind 1. Union	5	75.57	78.71	154.28
BFS Kind 1. Union	5	75.57	78.71	154.28
ES Kind 1. Ensemble	6	74.23	82.10	156.33
FFS Kind 1. Ensemble	5	75.18	82.10	157.27
BFS Kind 1. Ensemble	5	75.18	82.10	157.27
ES Kind 2. Union	10	75.73	80.32	156.05
FFS Kind 2. Union	2	72.65	77.10	149.75
BFS Kind 2. Union	2	72.65	77.10	149.75
ES Kind 2. Ensemble	6	76.05	79.19	155.24
FFS Kind 2. Ensemble	2	72.41	77.42	149.83
BFS Kind 2. Ensemble	2	72.41	77.42	149.83

5.6.3 AASA for ecstasy consumption by usage of several classifiers and ES, FFS, and BFS

Recall that required accuracy is 65%. The best FS model is KNN model based on ES, FFS, and BFS with usage of all attributes except A. This model has sensitivity 75.63% and specificity 75.75%. To find ensemble models for both kind of AASA we used outputs of several models for all sets as input feature to form the first layer of ensemble model and then we used linear regression to find linear weights for the second layer.

FFS based the first kind AAS

We apply FFS for ecstasy consumption using several classifiers. Recall that the best FS model is KNN model with all features exclude A and has sensitivity 75.63% and specificity 75.75%.

To select the first minimal set we apply FFS for each classifier. Table 5.13 presents found minimal sets for each classifier. Table 5.13 shows that set defined by GM is not a minimal one for several classifiers. Indeed, this set includes age, SS and

Table 5.13. First (minimal) feature set based on FFS for ecstasy usage by several classifiers, column '#' contains number of used features, 'X' means used input feature

Classifier	Model	#	Age	O	SS	Gndr	Sn (%)	Sp (%)	Sum(%)
LR		2	X	X			69.14	71.24	140.37
KNN	M_1	2	X	X			70.97	70.81	141.78
GM		3	X		X	X	67.64	77.87	145.51
LDA		2	X		X		68.43	77.50	145.93
PDFE		2	X	X			70.97	70.55	141.52
NB		2	X		X		65.38	77.69	143.07

Table 5.14. Second AAS of first kind based on FFS for ecstasy usage without age and O for several classifiers, column '#' contains number of used features, 'X' means used input feature

Classifier	Model	#	Edu	A	C	SS	Sn (%)	Sp (%)	Sum(%)
KNN		2			X	X	70.55	70.31	140.85
LDA		2		X		X	67.02	71.11	138.13
LR		2	X			X	68.78	68.98	137.76
PDFE	M_2	2	X			X	71.78	70.57	142.35

Gndr but minimal set for LDA and NB includes age and SS only. It means that GM selected set includes another minimal set. As a result, there are two minimal sets: age and O and age and SS. First set is minimal for three classifiers. The best accuracy is provided by KNN for set with age and O (this model is highlighted by yellow background in Table 5.13). It is the first AAS and is called M_1.

To find the first kind AAS for set M_1, we repeat FFS for all classifiers and for all feature exclude age and O. The found minimal sets for four classifiers are presented in Table 5.14. There is no minimal model for GM and NB and for feature set which does not contain age and O. The best accuracy is provided by PDFE for Edu and SS. This model is the second AAS (M_2) and highlighted by yellow background in Table 5.14.

To find the first kind AAS for set M_2, we repeat FFS for all classifiers and for all feature excluding age, O, Edu and SS. We found minimal sets for only KNN classifier. Table 5.15 shows that set defined by KNN is a minimal one. There is no minimal models for LR, GM, LDA, PDFE and NB and for feature set which does not contain age, O, Edu, and SS. The KNN model is the third AAS (M_3) and highlighted by yellow background in Table 5.15.

Table 5.15. Third AAS of first kind based on FFS for ecstasy usage without age, SS, O, and Edu for several classifiers, column '#' contains number of used features, 'X' means used input feature

Classifier	Model	#	C	Imp	Sn (%)	Sp (%)	Sum(%)
KNN	M_3	2	X	X	65.26	65.11	130.37

Table 5.16. Found FFS minimal sets for features without age for each classifier for ecstasy consumption, column '#' contains number of used features, 'X' means used input feature; non-minimal set is highlighted by blue background

Classifier	Model	#	A	O	SS	Imp	Gndr	Sn (%)	Sp (%)	Sum(%)
LR	M_2	2		X	X			70.44	67.90	138.34
LDA	M_3	2	X		X			71.11	67.02	138.13
KNN	M_4	2		X	X			70.19	70.17	140.37
PDFE	M_5	2		X	X			70.19	70.17	140.37
NB	M_6	4		X	X	X	X	72.40	65.91	138.31

Search of minimal sets for attributes N, E, A, and Gndr find nothing for all classifiers (all found models have inappropriate accuracy). As a result we have three sets which are AAS for each other: M_1, M_2 and M_3. Union of these sets contain six features. PDFE is the best classifier for this set. The PDFE union model has sensitivity 74.57% and specificity 74.96%. The ensemble model for these AAS has sensitivity 74.69% and specificity 74.83%.

FFS based the second kind AAS

Table 5.13 shows that the minimal set is M_1 and contains age and O. To find AAS of the second kind it is necessary to perform next procedure. For each classifier Cl_i and each element $e \in M_1$ we search minimal set $A_{ie} = M(F \setminus \{e\})$. Then we consider all sets A_{ie} and remove all non-minimal sets. Last step is important because two different classifiers can find two sets which are minimal for used classifiers but one of the set is part of the second set. Table 5.16 demonstrates such case: set defined by NB completely contains minimal set defined by LR, KNN, and PDFE. Table 5.16 lists all minimal sets which do not contain age and are selected by FFS for all classifiers.

Table 5.17 lists all minimal sets which do not contain O and are selected by FFS

Table 5.17. Found FFS minimal sets for features without O for each classifier for ecstasy consumption; column '#' contains number of used features, 'X' means used input feature; non-minimal set is highlighted by blue background

Classifier	Model	#	Age	SS	Gndr	Sn (%)	Sp (%)	Sum(%)
LR	M_7	2	X	X		74.70	70.64	145.34
LDA	M_8	2	X	X		77.50	68.43	145.93
KNN	M_9	2	X	X		71.87	71.64	143.51
PDFE	M_{10}	2	X	X		72.44	72.40	144.84
NB	M_{11}	2	X	X		77.69	65.38	143.07
GM	M_{12}	3	X	X	X	77.87	67.64	145.51

Table 5.18. First (minimal) feature set based on FFS for ecstasy usage by several classifiers, column '#' contains number of used features, 'X' means used input feature

Classifier	Model	#	Age	SS	Sn (%)	Sp (%)	Sum(%)
LR		2	X	X	70.64	74.70	145.34
KNN		2	X	X	71.87	71.64	143.51
GM		2	X	X	65.38	77.69	143.07
LDA		2	X	X	68.43	77.50	145.93
PDFE	M_1	2	X	X	72.44	72.40	144.84
NB		2	X	X	65.38	77.69	143.07

for all classifiers.

There are four models in Table 5.16 and five models in Table 5.17. However there are only two different minimal sets in Table 5.16 and only one minimal set in Table 5.17. The union model contains four attributes (age, O, A and SS). The best model for this feature set is PDFE and has sensitivity 73.64% and specificity 73.81%. The ensemble model for these AAS has sensitivity 74.17% and specificity 74.25%.

BFS based the first kind AAS

The best FS model is KNN model with usage of all attributes except A. This model has sensitivity 75.63% and specificity 75.75%. To select the first minimal set we apply BFS for each classifier. Table 5.18 presents found minimal sets for each classifier. There is one minimal set for each classifier which is age and SS. The best accuracy is provided by PDFE (this model is highlighted by yellow background in Table 5.18). It is the first AAS and is called M_1.

Table 5.19. First alternative of first kind based on BFS for ecstasy usage without age and SS by several classifiers, 'X' means used input feature

Classifier	Model	#	Edu	O	C	Imp	Sn (%)	Sp (%)	Sum(%)
KNN	M_2	2	X	X			68.08	67.91	135.99
LDA		4	X	X	X	X	65.70	67.91	133.61
LR		2	X	X			65.26	67.51	132.77
PDFE		2		X	X		67.64	67.51	135.15

To find the first kind AAS for set M_1, we repeat BFS for feature sets for all classifiers which does not contain age and SS. The minimal set M_2 is highlighted by the yellow background is the first AAS of set M_1 selected by KNN. Table 5.19 shows that set defined by LDA is not minimal one for several classifiers. LDA selected set includes another minimal set.

We do not have AAS for M_1 and M_2. We do not find set for any classifiers does not contain sets M_1 and M_2, and satisfy restriction to have at least 65% of sensitivity and specificity.

As a result we have two sets which are AAS for each other: M_1 and M_2. Union of this sets contain four features. PDFE is the best classifier for this set. This union model has sensitivity 74.17% and specificity 74.43%. The ensemble model for these two AASA has sensitivity 74.70% and specificity 74.52%.

BFS based the second kind AAS

The minimal set is M_1 and contains age and SS. To find AAS of the second kind it is necessary to repeat BFS for feature set by several classifiers without age and without SS separately. The result to find minimal set for all attribute exclude age by several classifiers are presented in Table 5.20. LDA and NB are not minimal one for several classifiers. GM cannot find minimal model with appropriate accuracy.

Found BFS minimal set for all attribute exclude SS by several classifiers are presented in Table 5.21. GM and NB is not minimal model for several classifiers.

The union model contain five attributes (age, O, SS, Edu, and C) and selected by

Table 5.20. Found BFS minimal sets for features without age for each classifier for ecstasy consumption, column '#' contains number of used features, 'X' means used input feature; non-minimal set is highlighted by blue background

Classifier	Model	#	Edu	C	O	Imp	SS	Gndr	Sn (%)	Sp (%)	Sum(%)
LR	M_2	2		X			X		69.64	68.52	138.16
KNN	M_3	2			X		X		70.19	70.17	140.37
NB	M_4	4		X	X	X	X		72.40	65.91	138.31
LDA	M_5	3		X			X	X	70.31	70.11	140.41
PDFE	M_6	2	X				X		71.78	70.57	142.35

Table 5.21. Found BFS minimal sets for features without SS for each classifier for ecstasy consumption, column '#' contains number of used features, 'X' means used input feature; non-minimal set is highlighted by blue background

Classifier	Model	#	Age	O	Edu	Gndr	Imp	C	A	Sn (%)	Sp (%)	Sum(%)
LR	M_7	2	X	X						71.24	69.14	140.37
LDA	M_8	2	X		X					71.78	67.11	138.89
KNN	M_9	2	X	X						70.97	70.81	141.78
PDFE	M_{10}	2	X	X						70.97	70.55	141.52
GM	M_{11}	4	X	X			X		X	77.60	66.71	144.31
NB	M_{12}	5	X	X		X	X	X		78.40	65.65	144.04

KNN and has sensitivity 74.83% and specificity 74.60%. The ensemble model for these AAS has sensitivity 74.70% and specificity 74.60%.

ES based the first kind AAS

The best ES model is selected by KNN and has sensitivity 75.63% and specificity 75.75%. List of all minimal sets is presented in Table D.5. There are three AAS of the first kind: the first is M_1 and includes age and SS, the second is M_{15} and includes Edu and O the third is M_{24} and contains C and Imp The best union model contains six features: age, Edu, O, SS, Imp, and C. PDFE is the best classifier for this set. The PDFE union model has sensitivity 74.57% and specificity 74.96%. The ensemble model for these three AAS has sensitivity 75.10% and specificity 75.04%.

ES based the second kind AAS

Table D.5 shows that the minimal set is M_1 selected by PDFE and contains age and SS. All other 28 sets are the second kind AAS for M_1 in accordance with definition. Union of these sets contains all features. The KNN union model is the best union model and has sensitivity 74.57% and specificity 74.25%.

Since we can use all 28 sets as the second kind AAS, we used outputs of several models for all sets as input feature to form ensemble model and we used linear regression to find weights. We should notice that we have 29 different sets for several classifiers (we have 100 models as input feature to find ensemble models. It means some sets are provided by more than one classifiers (see Table D.5)). Then we apply the FS methods BFS and FFS to select second kind of AAS. The ES is too expansive for 100 features. The best model selected by linear regression is found based on FFS with all features. This model has sensitivity 81.22% and specificity 81.89%.

Comparison of optimal models for ES, FFS, and BFS and all AASA models on base of ES, FFS and BFS for ecstacy use

The best FS models by several classifiers for ES, FFS, and BFS and all AASA results for ecstasy consumption are presented in Table 5.22. To show relation between minimal model and AAS based models for several classifiers, we include minimal models for each method in Table 5.22.

The AASA by using several classifiers could build model with higher accuracy than the best models based on FS methods: ES, FFS, and BFS. The best AASA model is ensemble model of the second kind for ES and has sensitivity 81.22% and specificity 81.89%. This model is selected by all classifiers with the first layer and linear regression used to find weights for the second layer and contains all features. This model has essentially better accuracy than best FS models.

Table 5.22. The results of FS models based on ES, FFS, and BFS and AASA results for ecstasy consumption by several classifier, column '#' contains number of used features, LW is linear regression used to find weights.

Model	#	Classifiers	Sn (%)	Sp (%)	Sum (%)
Best ES	9	KNN	75.63	75.75	151.38
Best FFS	9	KNN	75.63	75.75	151.38
Best BFS	9	KNN	75.63	75.75	151.38
ES Minimal	2	PDFE	72.44	72.40	144.84
FFS Minimal	2	KNN	70.97	70.81	141.78
BFS Minimal	2	PDFE	72.44	72.40	144.84
ES Kind 1. Union	6	PDFE	74.57	74.96	149.52
FFS Kind 1. Union	6	PDFE	74.57	74.96	149.52
BFS Kind 1. Union	4	PDFE	74.17	74.43	148.60
ES Kind 1. Ensemble	6	LW (LR, KNN, PDFE, LDA, NB, GM)	75.10	75.04	150.14
FFS Kind 1. Ensemble	6	LW (LR, KNN, PDFE)	74.69	74.83	149.53
BFS Kind 1. Ensemble	4	LW (LR, KNN, PDFE, LDA, NB, GM)	74.70	74.52	149.22
ES Kind 2. Union	10	KNN	74.57	74.25	148.82
FFS Kind 2. Union	4	PDFE	73.64	73.81	147.45
BFS Kind 2. Union	5	KNN	74.83	74.60	149.44
ES Kind 2. Ensemble	10	LW (LR, KNN, PDFE, LDA, NB, GM)	81.22	81.89	163.11
FFS Kind 2. Ensemble	3	LW (LR, KNN, PDFE, LDA, NB)	74.17	74.25	148.42
BFS Kind 2. Ensemble	2	LW (LR, KNN, PDFE, LDA, NB, GM)	74.70	74.60	149.30

5.6.4 AASA for cannabis consumption by usage of several classifiers and ES, FFS, and BFS

The best FS model is found for ES and BFS which is selected by LDA and has sensitivity 78.89% and specificity 79.84%. This model uses five attributes: age, Edu, O, A, and C. The best FS model based on FFS is LDA and has sensitivity 78.39% and specificity 79.29%. This model uses eight attributes: Age, Edu, E, A, C, Imp, and SS. The best models by several classifiers for ES, FFS and BFS and all results of AASA are presented in Table 5.23. There are 14 minimal feature sets with the best accuracy of several classification models. These minimal feature

Table 5.23. The results of FS models based on ES, FFS, and BFS and AASA results for cannabis consumption by several classifier, column '#' contains number of used features, LW is linear regression used to find weights.

Model	#	Classifiers	Sn (%)	Sp (%)	Sum (%)
Best ES	5	LDA	78.89	79.84	158.73
Best FFS	8	LDA	78.39	79.29	157.68
Best BFS	5	LDA	78.89	79.84	158.73
ES Minimal	2	PDFE	68.93	68.87	137.80
FFS Minimal	1	LR, KNN, LDA, PDFE	73.12	67.90	141.03
BFS Minimal	1	LR, KNN, LDA, PDFE	73.12	67.90	141.03
ES Kind 1. Union	7	GM	77.95	78.07	156.01
FFS Kind 1. Union	7	GM	77.95	78.07	156.01
BFS Kind 1. Union	7	GM	77.95	78.07	156.01
ES Kind 1. Ensemble	7	LW(LR, KNN, LDA, PDFE, GM)	81.26	80.97	162.23
FFS Kind 1. Ensemble	2	LW(LR, KNN, LDA, PDFE)	79.53	79.19	158.72
BFS Kind 1. Ensemble	2	LW(LR, KNN, LDA, PDFE)	79.53	79.19	158.72
ES Kind 2. Union	10	KNN	77.79	78.39	156.17
FFS Kind 2. Union	7	GM	77.95	78.07	156.01
BFS Kind 2. Union	7	GM	77.95	78.07	156.01
ES Kind 2. Ensemble	9	LW(LR, KNN, LDA, PDFE, GM)	81.29	81.34	162.63
FFS Kind 2. Ensemble	7	LW(LR, KNN, LDA, PDFE, GM, NB)	80.24	80.16	160.40
BFS Kind 2. Ensemble	7	LW(LR, KNN, LDA, PDFE, GM, NB)	79.92	79.84	159.76

sets are presented in Table D.4 in Appendix.

The best AASA model is ensemble model of the first kind for ES and second kind for ES, BFS, FFS. The ensemble model for the first kind for ES contains seven features and has sensitivity 81.26% and specificity 80.96%. The ensemble model for the second kind for ES contains nine features and has sensitivity 81.29% and specificity 81.34%. The ensemble model for the second kind for FFS and BFS contains seven features. For the ensemble models for the second layer we used linear regression to find weights.

We can see that applying the AASA with several classifiers to one data mining

problem allows us to build better accuracy models than the using AASA with one classifier to the same problem. For instance, for cannabis consumption using several classifier we found the ensemble model for the second kind based on ES is much better than the ensemble model for the second kind AAS based on one classifier (see Table 5.12 and Table 5.23). It is true for ecstasy usage as well.

5.7 AASA for USA president elections dataset

The second database is prediction of USA president elections (politics, [4]). This database contains 31 instances and 12 Boolean features. This dataset describe results of elections of the USA president for the period from 1860 to 1980. In each election there are two basic opponents: the aspirant of the party currently in power (P-party) and the aspirant of the opposition party (O-party). These 12 features are answers for questions about the political, economic, social conditions of the country, and candidates themselves. This data set described in [4] and [169]. List of questions is presented in Appendix D.2.

Recall that required accuracy for this data set is 75%. We have to use one or several classifiers to evaluate accuracy.

5.7.1 AASA for USA president elections by usage LR and ES, FFS, and BFS

The best FS model based on ES and BFS is selected by LR and have sensitivity 100% and specificity 100%. This model used five attributes: Q3, Q4, Q7, Q8, and Q12. The best FS model based on FFS is selected by LR and have the full accuracy, but used seven attributes instead of five (see Table 5.24).

All the best model and AASA results for president elections are shown in Table 5.24. The ES minimal model by using LR contains Q3, Q5, Q8, and Q9 features. This model provides high accuracy for prediction of the USA president

Table 5.24. The results of LR models selected by ES, FFS, and BFS and AASA results for president elections, column '#' contains number of used features, LW is linear regression used to find weights.

Model	#	Sn (%)	Sp (%)	Sum (%)
Best ES	5	100	100	200
Best FFS	7	100	100	200
Best BFS	5	100	100	200
ES Minimal	4	92	89	181
FFS Minimal	1	77	94	171
BFS Minimal	1	77	94	171
ES Kind1. Union	9	77	94	171
FFS Kind1. Union	9	77	100	177
BFS Kind1. Union	8	85	100	185
ES Kind 1. Ensemble	9	100	94	194
FFS Kind 1. Ensemble	9	100	100	200
BFS Kind 1. Ensemble	8	100	94	194
ES Kind 2. Union	12	92	100	192
FFS Kind 2. Union	7	85	94	179
BFS Kind 2. Union	3	77	72	149
ES Kind 2. Ensemble	8	100	100	200
FFS Kind 2. Ensemble	7	100	100	200
BFS Kind 2. Ensemble	3	77	94	171

elections. This minimal set has sensitivity 92% and specificity 89%.

The best AASA result is the FFS and ES ensemble model for the second kind and FFS ensemble model for the first kind of AASA with different number of input features. These models have full accuracy.

Graphs of accuracy versus number of used features for different FS methods are presented in Figure 5.8. We can see that almost all AASA results have approximately the same accuracy.

Figure 5.9 shows the sets of features which are used in the best ES model, in the best ensemble and union models for the first kind of AAS and for the second kind of AAS.

5.7.2 AASA for USA president elections by usage of several classifiers and ES, FFS, and BFS

Since president elections problem has 12 attributes only and ES can be completed for reasonable time.The best FS model is selected by LR and has sensitivity 100%

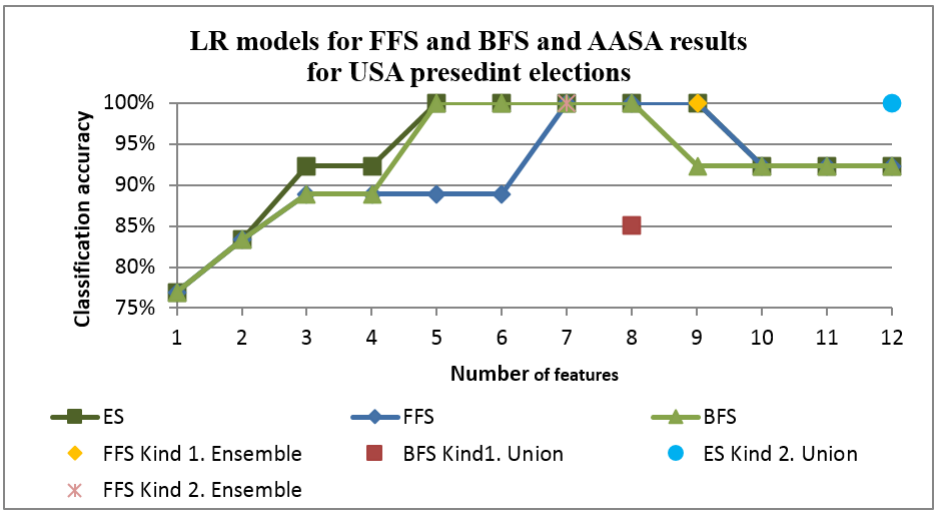

Figure 5.8. Comparisons of LR models selected by ES, FFS, and BFS and AAS results for USA president elections.

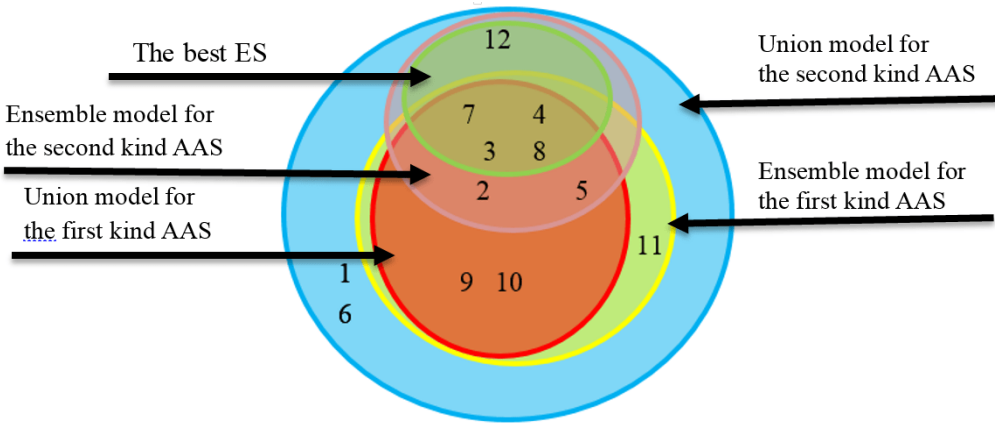

Figure 5.9. Venn diagram for features which are used for the best ES model, the first kind ensemble and union models and the second kind ensemble and union models for USA president elections.

and specificity 100%. It means LR is the optimal classifier among all used classifiers. The required accuracy is 75%.

For president elections there are 86 minimal feature sets selected by several classifiers (i.e 81 different sets part of which is selected by several classifiers). These minimal feature sets are presented in Table D.9 in Appendix.

All results for president elections are shown in Table 5.25. The best AASA result is the ensemble model for the second kind of AAS with eight input features for ES and BFS. ES is too expensive with 86 input features. We applied FFS for secondary model selection. Another best AASA result is the ensemble model for the first

Table 5.25. The results of FS models based on ES, FFS, and BFS and AASA results for USA president elections by several classifier, column '#' contains number of used features, LW is linear regression used to find weights.

Model	Classifiers	#	Sn (%)	Sp (%)	Sum (%)
Best ES	LR	5	100	100	200
Best FFS	LR	7	100	100	200
Best BFS	LR	5	100	100	200
ES Minimal	LR, KNN, NB	4	85	83	168
FFS Minimal	LR, LDA, KNN, NB &GM	1	77	94	171
BFS Minimal	LR, LDA, KNN, NB &GM	1	77	94	171
ES Kind 1. Union	KNN	11	92	94	187
FFS Kind 1. Union	LR	12	92	100	192
BFS Kind 1. Union	LR	11	92	100	192
ES Kind 1. Ensemble	LW(LR, GM, NB)	5	100	94	194
FFS Kind 1. Ensemble	LW(LR,KNN, LDA, GM, NB)	12	100	100	200
BFS Kind 1. Ensemble	LW(LR,KNN, LDA, GM, NB)	7	100	94	194
ES Kind 2. Union	LR	12	92	100	192
FFS Kind 2. Union	LR	12	92	100	192
BFS Kind 2. Union	LR	12	92	100	192
ES Kind 2. Ensemble	LW(LR, GM, NB)	8	100	100	200
FFS Kind 2. Ensemble	LW(LR,KNN, LDA, GM, NB)	12	100	100	200
BFS Kind 2. Ensemble	LW(LR,KNN, LDA, GM, NB)	8	100	100	200

kind for FFS with all features. We achieved those results when we use output result for the first layer as input feature for the second layer and we used linear regression to find weights.

5.8 AASA for Breast cancer dataset

The third database which is used to AASA applications is available dataset of find needle aspirates (FNA) of breast cancer. The FNAs used to develop the diagnostic system [5]. This dataset contains 569 samples of FNA of breast cancer, including 212 positive specimens with cancer (malignancy) and 357 negative specimens with fibrocystic breast masses (benign). All of the samples were confirmed by us-

Table 5.26. Numbers of features of Breast cancer dataset

Measurement	Mean	SE	Worst
radius (mean distance from center to border)	1	11	21
texture (standard deviation of gray-scale values)	2	12	22
perimeter	3	13	23
area	4	14	24
smoothness (local variation in radius lengths)	5	15	25
compactness (perimeter^2 / area - 1.0)	6	16	26
concavity (severity of concave portions of the contour)	7	17	27
concave points (number of concave portions of the contour)	8	18	28
symmetry	9	19	29
fractal dimension ("coastline approximation" - 1)	10	20	30

ing a computer based system for diagnosing breast FNAs [5]. Each sample contains ten features of the cell. The ten features which computed for each nucleus are: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The system uses computer vision process to analyse nuclear size, shape, and texture attributes and organizes features using an inductive technique created on linear programming [174]. These features are used to differentiate between benign and malignant breast cytology.

The mean value, worst (mean of the three largest value), and Standard Error (SE) of each feature are computed for each image. As a result, the database contains 30 features [5, 61]. The observations and details of the cellular features were all created by consultant pathology [5,61]. We use for referencing a numbers of features in accordance with Table 5.26.

5.8.1 AASA for breast cancer by usage LR and FFS and BFS

FS is a significant steps in breast cancer detection and classification. Let us consider AASA for breast cancer diagnosis. Since the number of nuclear feature sets is 30 the ES is unacceptable for this problem. FFS and BFS techniques are being considered for selecting feature subset with high accuracy. Let us find AAS of feature sets. The best BFS model selected by LR has sensitivity 98.60% and specificity 96.70% and use 14 attributes: 1, 5, 6, 7, 8, 13, 14, 18, 20, 22, 23, 28, 29, and 30. The best FFS model selected by LR has sensitivity 98.60% and specificity 97.64%

Table 5.27. The results of FS models based on ES, FFS, and BFS and AASA results for Breast Cancer by several classifier, column '#' contains number of used features, LW is linear regression used to find weights.

Model	#	Sn (%)	Sp (%)	Sum (%)
Best FFS	9	98.60	97.64	196.24
Best BFS	14	98.60	96.70	195.30
FFS Minimal	3	95.80	96.23	192.03
BFS Minimal	4	95.28	96.08	191.36
FFS Kind1. Union	19	96.92	94.34	191.26
BFS Kind1. Union	23	96.92	94.81	191.73
FFS Kind 1. Ensemble	19	98.04	98.11	196.15
BFS Kind 1. Ensemble	23	98.04	97.17	195.21
FFS Kind 2. Union	8	98.32	95.76	194.07
BFS Kind 2. Union	12	97.76	97.17	194.93
FFS Kind 2. Ensemble	8	98.60	97.17	195.77
BFS Kind 2. Ensemble	12	98.04	97.64	195.68

and use nine attributes: 1, 3, 8, 15, 20, 22, 24, 26, and 30. The accuracy of visual diagnosis breast FNA is reported in several of the studies is above 90% [61]. In 37 series previous research stated by [175] and more than 25 series study (e.g. see [5, 61]) with a whole of 23,741 satisfactory breast FNA shows that the total accuracy is 94.3%. In current study to implement AAS of FS let us consider 95% as required accuracy.

The best LR classification results for breast cancer diagnosis are shown in Table 5.27. The minimal feature sets based on FFS and BFS are includet into Table 5.27. The minimal set based on FFS contains three features: mean concave points, worst area, and worst texture. This minimal set has sensitivity 95.80% and specificity 96.23%. The minimal set based on BFS contains four features. This minimal set has sensitivity 95.28% and specificity 96.08%. For this data base the minimal sets are better than the union model created on base first kind AAS.

The best AASA result is the ensemble model for the first kind based on FFS. This model is much better than the best FFS model but contains 19 features instead of nine. This LR model has sensitivity 98.04% and specificity 98.11%.

Another best AASA result is BFS ensemble model for second kind AAS which is much better than best BFS model. This model has sensitivity 98.04% and speci-

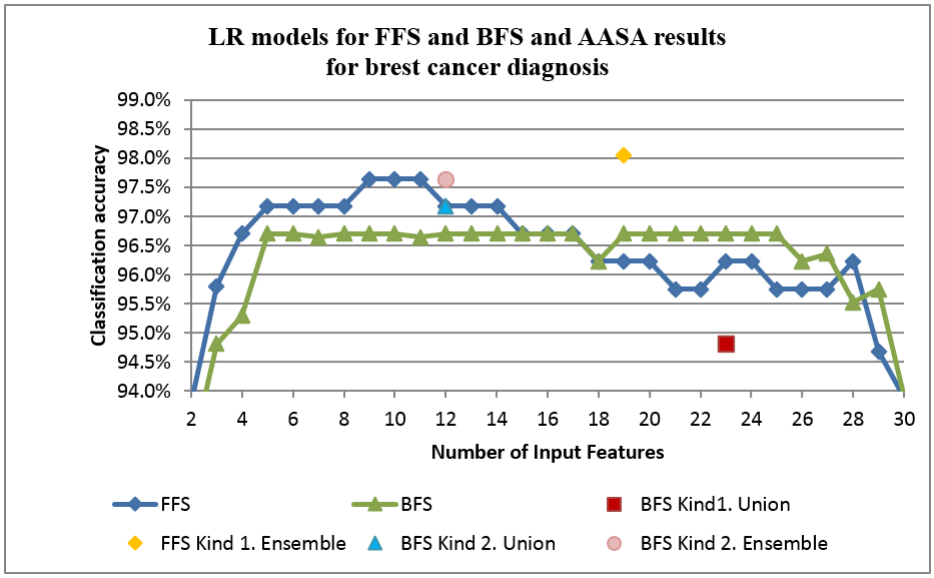

Figure 5.10. Comparisons of LR models selected by FFS and BFS and AAS results for breast cancer diagnose.

ficity is 97.64%. These results are better than the 97.3% accuracy based on the best single-plane diagnostic classifier based on features mean texture, the worst area, and the worst smoothness obtained by [61]. The results show that the best minimal model based on FFS contains three of the 30 features: mean concave points, worst area, and worst texture used by LR. This model has sensitivity 95.80% and specificity 96.23%. This result is considerably better than 89% accuracy based on individual cell analysis obtained by Hutchinson et al [62], and also considerable better than LR cross-validated classification accuracy is 96.2% obtained by Wolberg et al [61] but for three different features: worse radius, worse texture, and worse concave points.

Therefore, AASA improved accuracy over the best diagnostic accuracy of breast cancer FNAs. Graphs of accuracy versus number of used features for FFS and BFS methods are presented in Figure 5.10.

We notice that ensemble models based on FFS and BFS are much better than the best FFS and best BFS models. Figure 5.11 shows Venn diagram to illustrate the relationship between input features which are used for the first kind ensemble and union models and the second kind ensemble and union models.

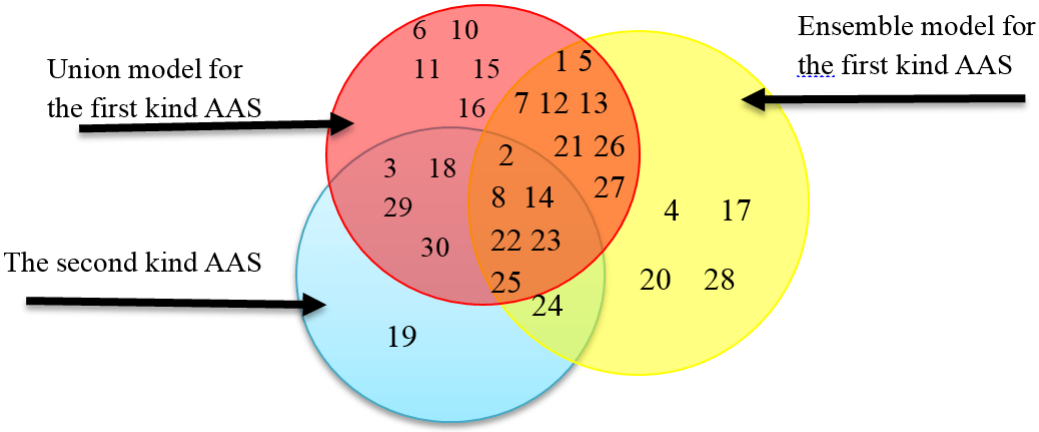

Figure 5.11. Venn diagram for features which are used for the first kind ensemble and union models and the second kind ensemble and union models for breast cancer diagnosis.

5.8.2 AASA for breast cancer by usage of several classifiers and FFS and BFS

The best FS model is selected by GM for FFS and has sensitivity 98.60% and specificity 97.64%. All results for are shown in Table 5.28. The best AASA result is the ensemble model for the first kind selected by KNN and LR for BFS with 23 attributes, the ensemble model for the second kind selected by LR for FFS with 14 attributes.

5.9 Discussion

In this chapter, we introduced a methodology of controlled multicollinearity, called AASA. The basic operation is the search of a minimal feature set with required accuracy. Minimal feature set can be found by ES, FFS, and BFS. We used six classifiers to select the best feature set with the highest accuracy.

We applied AASA for three datasets with different number of records, different dimensions and are taken from different application areas. The first database is drug consumption (psychology, [2, 3], section 'Database' of chapter 2). We create AASA models for three drugs: heroin, ecstasy, and cannabis. The second

Table 5.28. The results of FS models based on ES, FFS, and BFS and AASA results for Breast Cancer by several classifier, column '#' contains number of used features, LW is linear regression used to find weights.

Model	Classifier	#	Sn (%)	Sp (%)	Sum (%)
Best FFS	GM, LR	8	98.60	97.64	196.24
Best BFS	GM	8	97.64	97.20	194.84
FFS Minimal	LR	3	95.80	96.23	192.03
BFS Minimal	KNN	3	96.23	96.64	192.87
FFS Kind1. Union	LR	22	95.76	96.36	192.11
BFS Kind1. Union	LR	23	96.23	96.64	192.87
FFS Kind 1. Ensemble	LW(KNN &LR)	22	97.20	97.64	194.84
BFS Kind 1. Ensemble	LW(KNN &LR)	23	98.32	98.11	196.43
FFS Kind 2. Union	KNN	26	96.64	95.76	192.39
BFS Kind 2. Union	KNN	30	96.36	96.70	193.06
FFS Kind 2. Ensemble	LW(LR)	14	98.88	98.58	197.46
BFS Kind 2. Ensemble	LW(LR, NB, GM. KNN, LDA, PDFE)	30	98.32	96.23	194.55

database is prediction of USA president elections (politics, [4]). The third dataset is breast cancer diagnosis (medicine, [61]). Required accuracies for minimal set identification are defined as 65% for grag use dataset, 75% for USA president elections dataset and 95% for breast cancer dataset.

It was shown that for heroin consumption by LR, ensemble model for the first kind AAS and for the second kind AAS for ES are much better than the best FS models. The ensemble model for the first kind AAS uses eight features and has sensitivity 71.23% and specificity 72.15% while ensemble model based on FFS has the maximum sum of sensitivity and specificity. The ensemble model for the second kind AAS uses ten features and has sensitivity 71.50% and specificity 71.70% (see Table 5.10). For ecstasy consumption the best AASA result is ensemble model for the first kind for BFS with four features and ensemble model for the second kind for ES with nine features (see Table 5.11). However, we have to use several classifiers with AASA. For example, AASA results for cannabis consumption have a good results than the best FS by several classifiers. The best AASA result for cannabis consumption is ES ensemble model for the first kind of AAS and ES

ensemble model for the second kind of AAS.

For the USA president elections data set the results illustrated that four features of the 12 are enough to achieve high accuracy in prediction of the USA president elections. It was shows that the optimal minimal model based on ES by using LR contains four features. This result was reproduced by several classifiers. The best ES model contains five features and has sensitivity 100% and specificity 100%. This model is exactly contained in second kind AAS ensemble model FS based on FFS which contains seven features and has the same accuracy (see Table 5.24). The best AASA result by several classifiers is ensemble model for second kind AAS for ES and BFS contains eight features and has sensitivity 100% and specificity 100%.

For the third data set the results represented an improvement over the best diagnostic accuracy of breast cancer. The results show that the optimal minimal model based on FFS contains three of the 30-features: mean concave points, worst area, and worst texture which were used by the LR model. This model has sensitivity is 95.80% and specificity is 96.23%. This result is considerably better than 89% accuracy based on individual cell analysis obtained by [62] and LR cross validated classification accuracy is 96.2% obtained by [5] but for three other features: worse radius, worse texture, and worse concave points. The best AASA result by LR is ensemble model for first kind for FFS with 19 input features. This model has sensitivity 98.04% and specificity 98.11%.

The best AASA result by using several classifiers for breast cancer is the ensemble model for the first kind selected by KNN and LR for BFS with 23 attributes which has sensitivity 98.32% and specificity is 98.11% and the ensemble model for the second kind selected by LR for FFS with 14 attributes which has sensitivity 98.88% and specificity is 98.58%.

It means that AASA can create more accurate models. We can see that the ensemble models are usually much better than the best FS model.

CHAPTER 6

Conclusion and outlook

Results of data mining of drug consumption dataset are important as they examine the question of the relationship between drug use and personality traits and engage the challenge of untangling correlated personality traits (the FFM, impulsivity, and sensation seeking [181]), and clusters of substance misuse (the correlation pleiades). The work acknowledged the breadth of a common behaviour which may be transient and leave no impact, or may significantly harm an individual. We examined drug use behaviour comprehensively in terms of the many kinds of substances that may be used (from the legal and anodyne, to the deeply harmful), as well as the possibility of behavioural over-claiming. We built into this study the wide temporality of the behaviour indicative of the chronicity of behaviour and trends and fashions (e.g. the greater use of LSD in the 1960s and 1970s, the rise of ecstasy in the 1980s, some persons being one-off experimenters with recreational drugs, and others using recreational substances on a daily basis).

We defined substance use in terms of behaviour rather than legality, as legislation in the field is variable. This data were gathered before 'legal highs' emerged as a health concern [182] so we did not differentiate, for example, synthetic cannabinoids and cathinone-based stimulants; these substances have been since widely made illegal. We were nevertheless able to accurately classify users of these substances (reciprocally, this data were gathered before cannabis decriminalisa-

tion in parts of North America, but again, we were able to accurately classify cannabis users). We included control participants who had never used these substances, those who had used them in the distant past, up to and including persons who had used the drug in the past day, avoiding the procrustean data-gathering and classifying methods which may occlude an accurate picture of drug use behaviour and risk [183]. Such rich data and the complex methods used for analysis necessitated a large and substantial sample.

The study was a theoretical regarding the morality of the behaviour, and did not medicalise or pathologise participants, optimising engagement by persons with heterogeneous drug-use histories. This study used a rigorous range of data-mining methods beyond those typically used in studies examining the association of drug use and personality in the psychological and psychiatric literature, revealing that decision tree methods were most commonly effective for classifying drug users. We found that high N, low A, and low C are the most common personality correlates of drug use, these traits being sometimes seen in combination as an indication of higher-order stability and behavioural conformity, and, inverted, are associated with externalisation of distress [176–178]. Deviation from this rule (N$\Uparrow$, A$\Downarrow$, C$\Downarrow$) for some drugs is also interesting. LSD use correlates with high O and low C (and does not correlate significantly with high N, at least, for recent LSD users). High O is correlated with use of many drugs.

Low stability is also a marker of negative urgency [76] whereby persons act rashly when distressed. In this work we points to the importance of individuals acquiring emotional self-management skills anteceding distress as a means to reduce self-medicating drug-using behaviour, and the risk to health that injudicious or chronic drug use may cause.

The main topic of this thesis was to develop the classifiers to analyse the predictability of the classification problems. The main problem was in the search and validation of psychological predictors of consumption of different drugs. We used several algorithms in order to analyse the predictability of user/non-

user classification on the basis of psychological data: decision trees with several split criteria (information gain, Gini gain or DKM gain), random forests, *k*-nearest neighbours with distances: the Euclidean distance, the Fisher's transformed distance and the adaptive distance, linear discriminant analysis, Gaussian mixtures, probability density function estimation by radial basis functions, logistic regression and naïve Bayes approach were applied to predict the risk of drug consumptions.

In chapter 1, we explained the notion of drug use and which personality traits we used to analyse the predisposition to use of drugs. We also reviewed some previous pertinent results, describe the problem, and briefly outline the answer. In addition, we reviewed of feature selection problem and we presented how to use redundant and alternative input features for apply data set.

In chapter 2, we presented and described analysis of the database drug consumption with information of 1885 respondents, usage of 18 drugs, and the method of data collection. The four different definitions of drug users we used were based on recency of use: decade-based, year-based, month-based, or week-based definitions. The numbers of users of 18 different psychoactive substances for the four definitions of users in the database were presented. We presented several available databases which we used to AASA testing.

In chapter 3, we reviewed the used methods of data analysis, from elementary T-scores to non-linear Principal Component analysis (PCA), including polychoric correlation, nonlinear CatPCA (Categorical Principal Component Analysis), sparse PCA, method of principal variables, original 'double Kaiser selection' rule, *k*NN for various distances, DT with different split criteria (information gain, Gini gain or DKM gain), LDA, GM, PDFE by radial-basis functions, LR, NB approach, RF, and criteria for selecting the best method. We briefly identified how to use alternative attributes sets which can be useful to build a good predictor. We described the approaches to define weights of ensemble models: Nelder-Mead, Luus-Jaakola, pattern search, and linear regression.

In chapter 4, we examined the potential effect of NEO-FFI personality traits, impulsivity, sensation-seeking, and demographic data on drug consumption for different drugs and for different user/non-user separation of drug users. We illustrated association between personality profiles (i.e. NEO-FFI-R) and drug consumption. We revealed that the personality profiles are significantly associated with users and non-users groups of the 18 drugs. We pointed out that personality score of drug users of all 18 drugs is moderately high $(+)$ or neutral (0) on N and O, and moderately low $(-)$ on both the personality profile concerning A and C. The effect of the personality profile concerning E is drug specific. For example, for the decade-based user/non-user definition the E score is negatively correlated with consumption of crack, heroin, VSA, and methadone (E score is $(-)$ for their users). It is has no predictive value for other drugs for the decade-based classification (the E score for users is (0)), whereas in the year-, month-, and week-based classification problems all three possible values of E score are observed.

We evaluated the individual drug consumption risk separately, for each drug. We analysed interrelations between the individual drug consumption risks for different 18 drugs. We tested several types of classifiers for each drug for the decade-based user definition to predict the risk of drug consumption. For each drug, the most effective subset of input attributes was selected to provide the highest level of accuracy. LOOCV was used for all tests in this study. In this study we select the classification method which provides the maximal value of the least of sensitivity and specificity as the best one. If there is a tie on this basis, as there is in two cases, the method with maximal sum of the sensitivity and specificity is selected as the best. An exhaustive search was performed to select the most effective subset of input features, and data mining methods to classify users and non-users for each drug. The best results with sensitivity and specificity being greater than 75% were achieved for some drugs: cannabis, crack, ecstasy, legal highs, LSD, and VSA. Sensitivity and specificity greater than 70% were achieved for amphetamines, amyl nitrite, benzodiazepines, chocolate, caffeine, heroin, ketamine, methadone and nicotine. We considered four user/non-user separation:

decade-, year-, month-, and week- based classifications problems. Structure of correlations of different drug users for year- and decade-based problems are approximately the same. We found three group of drugs, each of group contains several drugs which were strongly correlated. That is the drug consumption had a 'modular structure'. The idea to unite correlated attributes into 'modules' called as *correlation pleiades* is popular in biology [43–45]. We evaluated the individual risk of drug consumption separately, for pleiad of drugs: *heroin, ecstasy, and benzodiazepines*. Users were defined for each correlation pleiad of drugs as users of any of the drug from the pleiad. We considered classification problem for drug pleiades for decade, year, month, and week-based user/non-user classification problems. The results of the classifiers evaluation were presented and the quality of classification was significantly high. We could see that for month-based problem of heroin pleiad consumption the best classifier was DT with five features and sensitivity 74.18% and specificity 74.11%. DT with seven attributes was the best classifier for year-based problem ecstasy pleiad consumption and has sensitivity 80.65% and specificity 80.72%. For a week-based problem of benzodiazepines pleiad consumption the best classifier was DT with five features and sensitivity 75.10% and specificity 75.76%.

The creation of classifiers provided the capability to evaluate the risk of drug consumption in relation to individuals. The risk maps is provided a useful tool for data visualization and the generation of hypotheses.

In chapter 5, we developed a methodology to crater controlled multicollinearity. This approach allows us selection of many types of alternative sets of input feature and usage all these alternative sets together. This technique we called AASA. It could find many different sets of relevant features each of which can be used to solve original problem separately. The concept of AAS was based on the notion of minimal feature subset. Minimal set can be found by one of the feature selection methods: ES, FFS, or BFS. We used either one or several classifiers to select minimal set. We aimed to obtain the best feature subset of the relevant attributes

from the original ones and the best classifier for each target feature. To select minimal set the required accuracy must be specified. LOOCV [120] was used to evaluate sensitivity and specificity. We defined two kinds of AAS. The first kind AAS for feature set *S* is the minimal set which does not contain any elements of S. The second kind AAS is the list of sets with two properties: (i) each set does not contain at least one element of set S and (ii) for each element of *S* there is at least one set which does not contain this element. We applied AASA to three different datasets. Accuracy restriction was: minimum of sensitivity and specificity had to be at least 65% for drug consumption dataset and 75% for US president elections dataset, and 95% for breast cancer dataset.

We defined five candidates to be the best model for each classification problem and for each FS method: (1) The best model among models which are tested by basic FS method (ES, FFS, or BFS). We called it the best FS model. The four candidate's model which comes from AASA are: (2) union model for the first kind AAS, (3) ensemble model for the first kind AAS, (4) union model for the second kind AAS, and (5) ensemble model for the second kind AAS. Then we selected the best model among these five candidates. We found that the best AASA model usually is much better than the best FS model. In most cases the best AASA model was ensemble model. For example, for heroin consumption ensemble model for the first kind AAS for ES is much better than the FS model. Moreover, the best FS model for ecstasy consumption contains seven features. This model has sensitivity 74.83% and specificity 74.52%. However, the best AASA model for ecstasy consumption was ensemble model for the first kind AAS for BFS. This ensemble model contains 4 feature instead of 7 and has better accuracy: sensitivity is 74.97% and specificity is 74.78%.

In this work, the results showed that for US president elections dataset 4 features of the 12 is enough to achieve success elections of the USA president. This result was very accurate in one or several classifiers. The results showed that the optimal minimal model based on ES by using LR contains four features. The best ES

model contains five features and has sensitivity 100% and specificity 100%. This optimal minimal ES model selected by three classifiers: NB, LR, and KNN. This model exactly containing in ensemble model for second kind AASA based on FFS which contains seven features with the same accuracy (see Figure 5.9).

For breast cancer problem we also demonstrated that ensemble model was much better than the best model. The best AASA result by using several classifiers is the ensemble model for the second kind selected by LR based on FFS with 14 attributes which has sensitivity 98.88% and specificity is 98.58% and the ensemble model of the first kind selected by KNN and LR based on BFS with 23 attributes which has sensitivity 98.32% and specificity is 98.11%.

Ensemble model for the first kind AAS for FFS is better than the best FFS model, but contains 19 features instead of 9. Accuracy of the ensemble model for the second kind AAS selected by LR and FFS with 14 attributes has accuracy which is much better than the 97.3% accuracy based on the best single-plane diagnostic classifier based on features mean texture, the worst area, and the worst smoothness which was obtained by Wolberg et al [61]. In addition, the best LR minimal model for FFS contains 3 of the 30 features: mean concave points, worst area, and worst texture. This model had sensitivity 95.80% and specificity 96.23%. This accuracy is considerably better than 89% accuracy of model based on individual cell analysis which was obtained by Hutchison et al. 1991 [62] and than 96.2% of the LR cross validated classification accuracy which is obtained by Wolberg et al [61] for other three features: worse radius, worse texture, and worse concave points.

We demonstrated that the AASA can significantly improve accuracy of model. We found that in AASA ensemble models are usually much better than the best FS method. AASA is the way how to use controlled multicollinearity or redundancy to create a better model.

APPENDIX A

Flowcharts of classification methods

The data analysis methods and procedures used in this work are represented by a system of flowcharts. The most flowcharts contain the legend in right top corner which explain the meaning of used variables. We illustrated flowcharts for the classifiers DT, kNN, and PDFE.

In all flowcharts the variables are highlighted by italic font. Flowcharts use the following kinds of blocks:

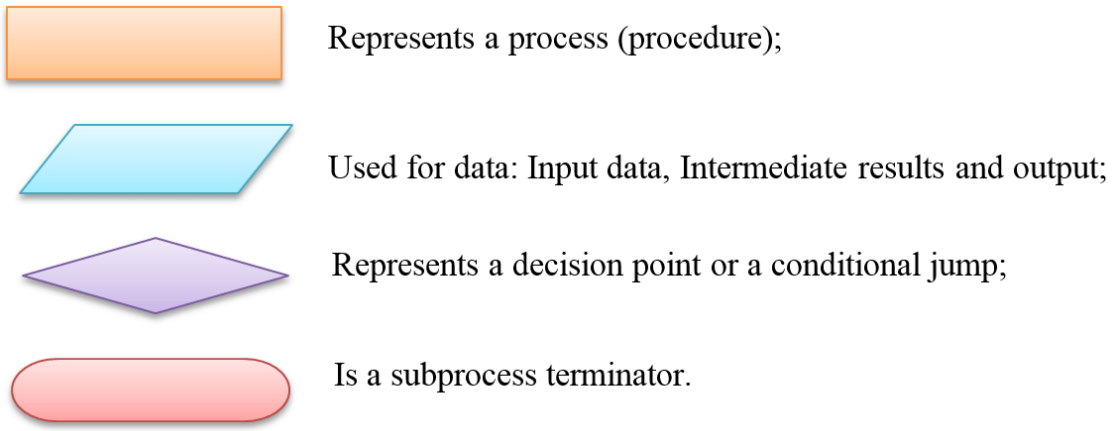

Figure A.1. Flowcharts blocks.

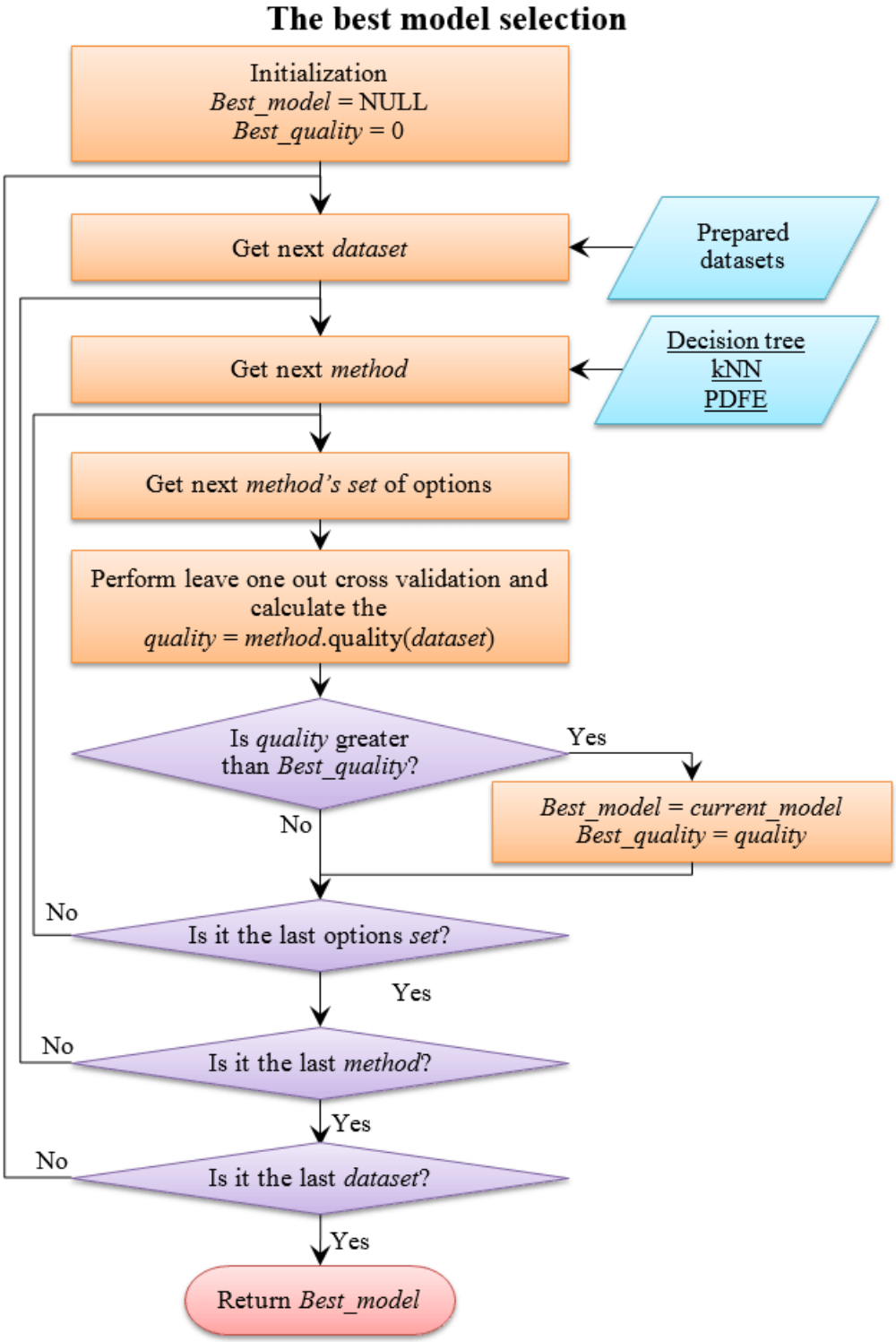

Figure A.2. The best model selection.

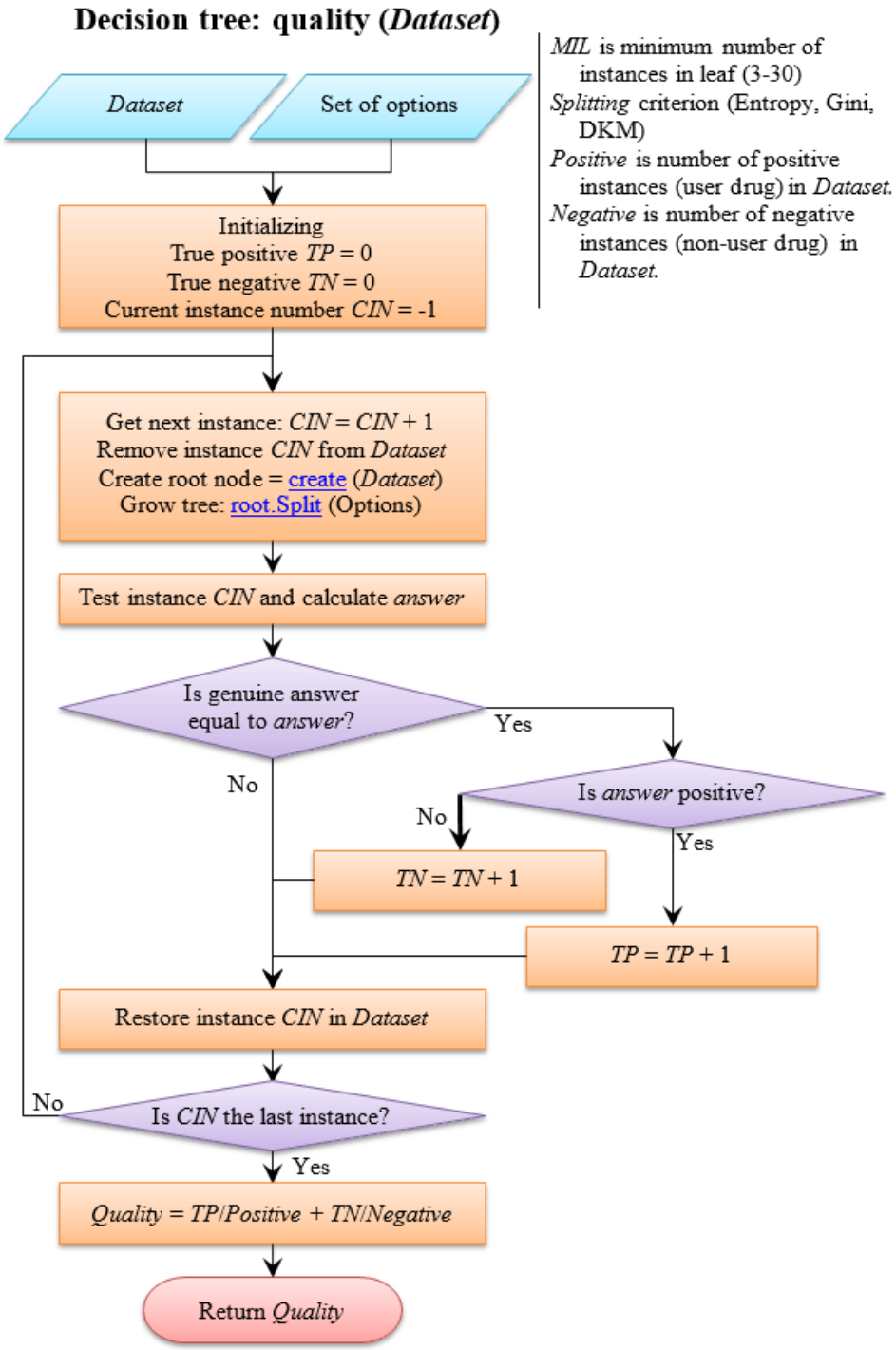

Figure A.3. Decision tree: quality (Dataset).

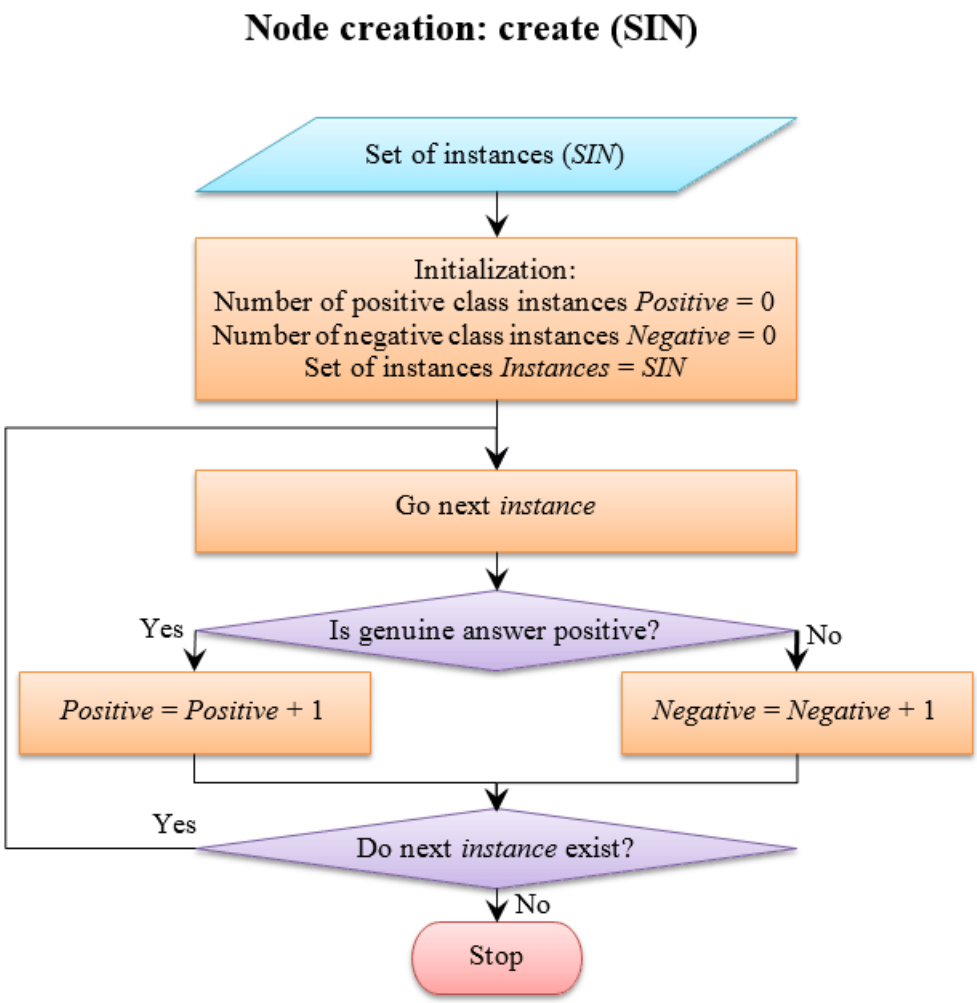

Figure A.4. Node creation: create (SIN).

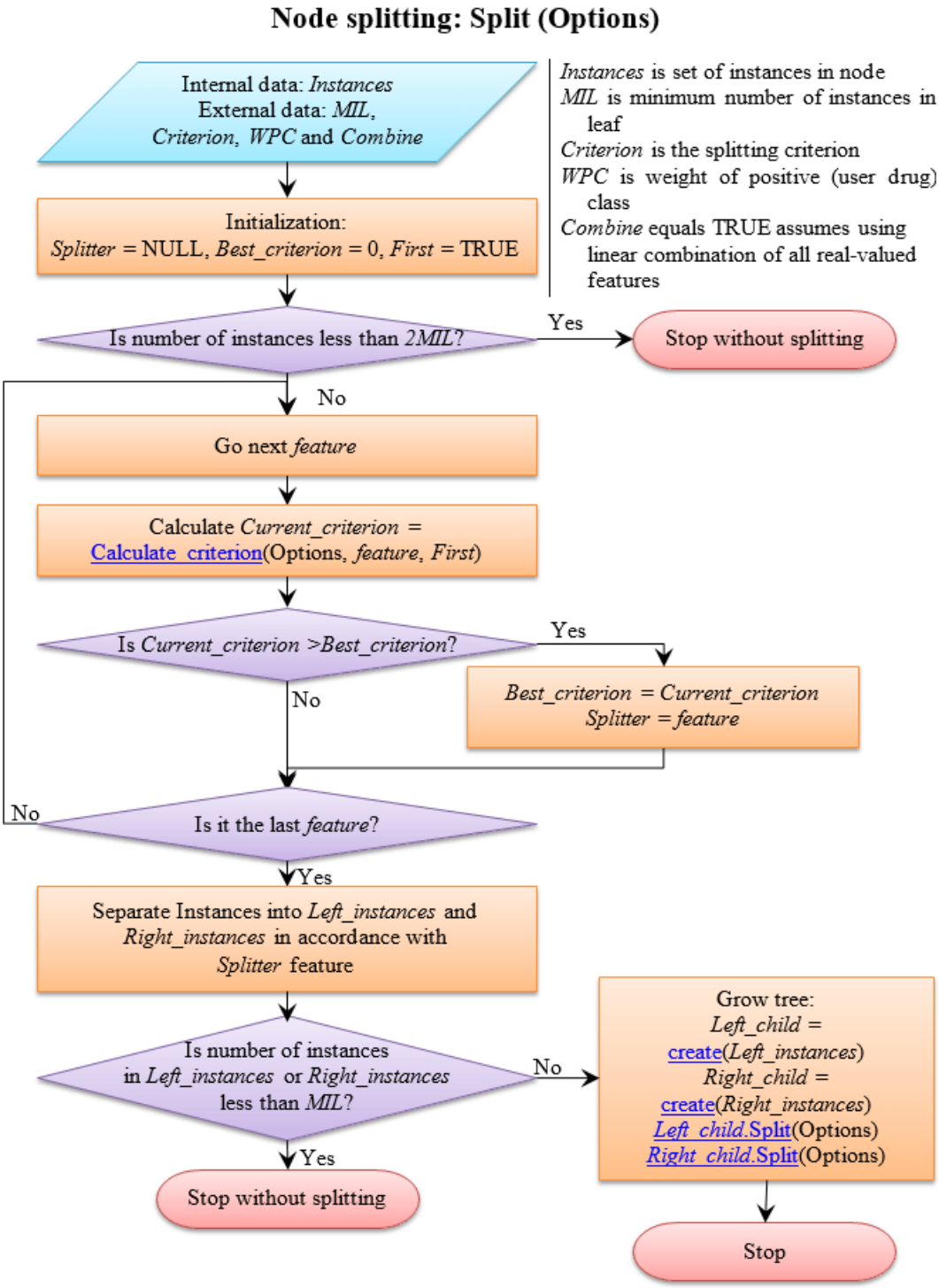

Figure A.5. Node splitting: Split (Options).

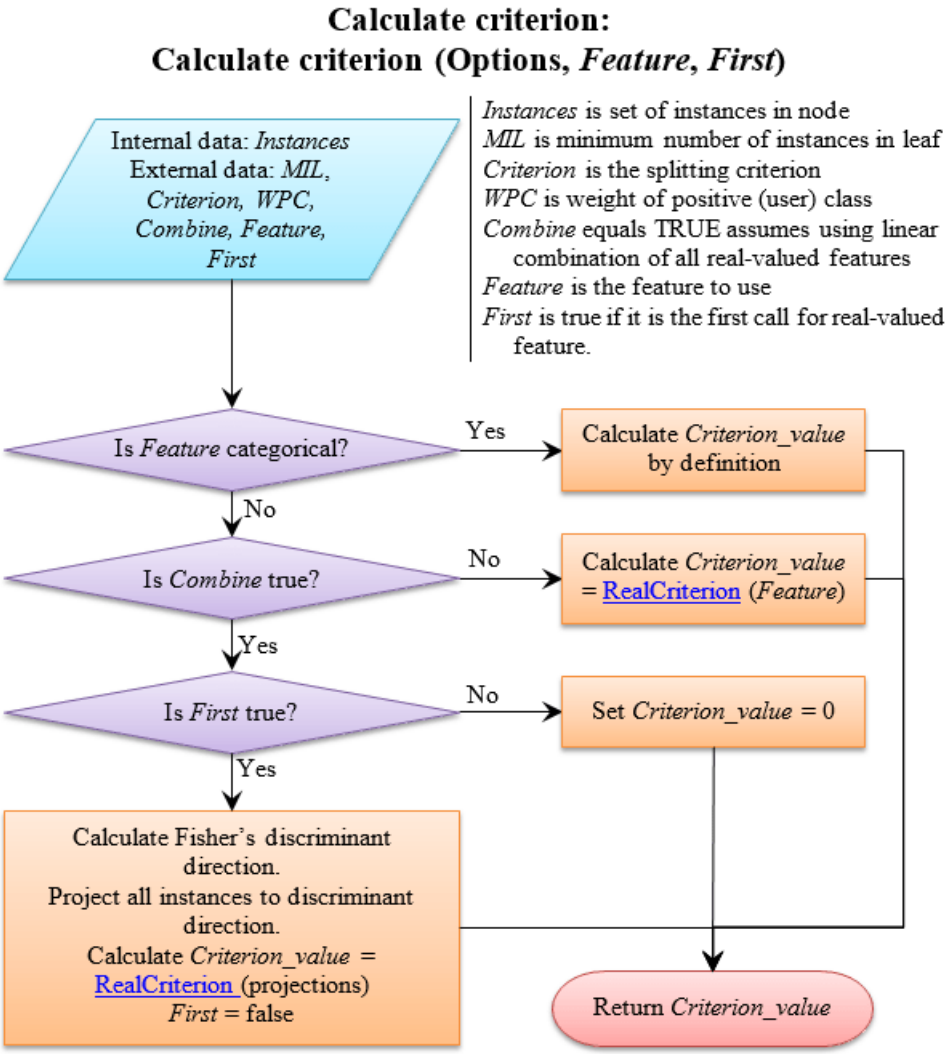

Figure A.6. Calculate criterion: Calculate criterion (Options, Feature, First).

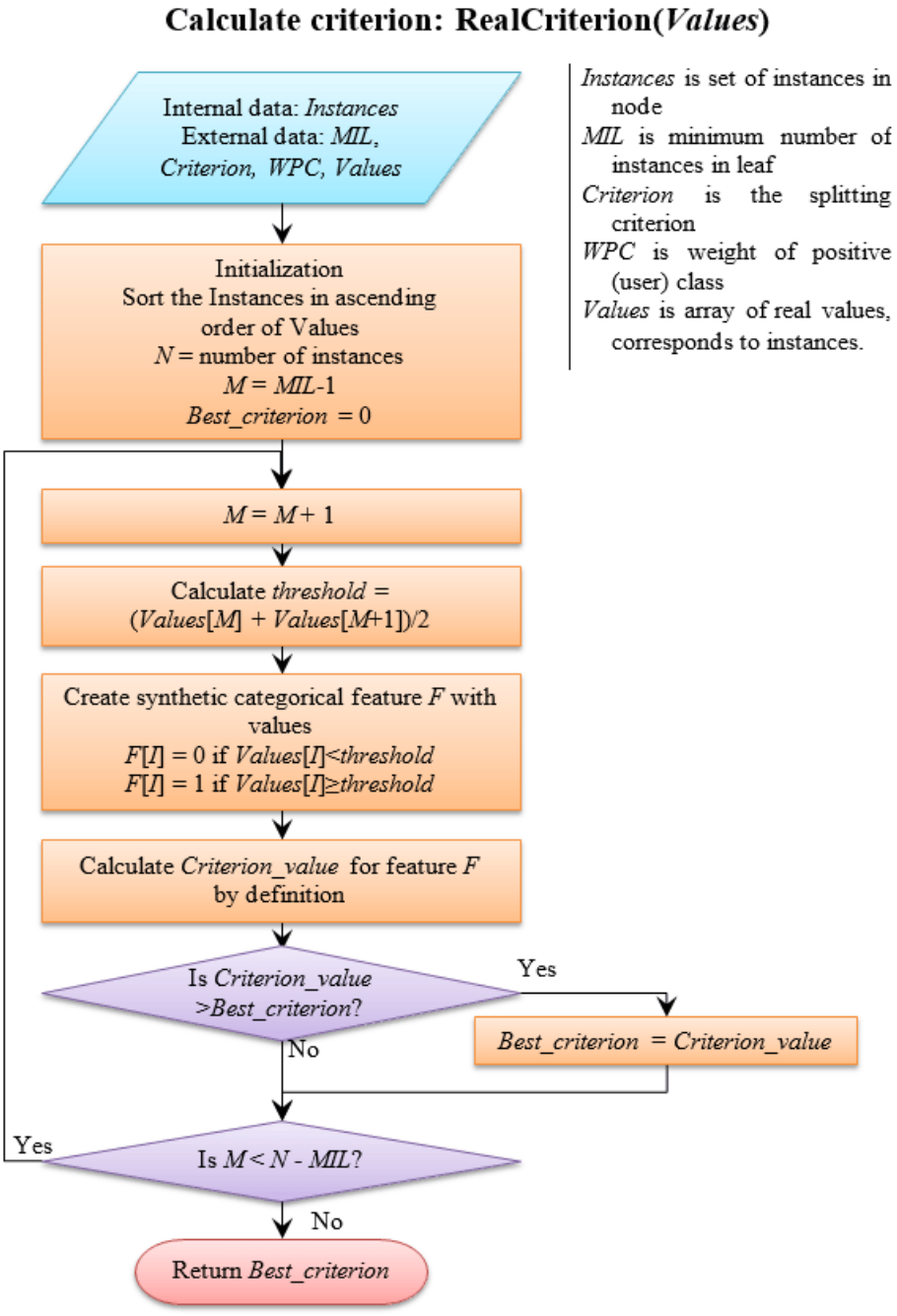

Figure A.7. Calculate criterion: Real Criterion(Values).

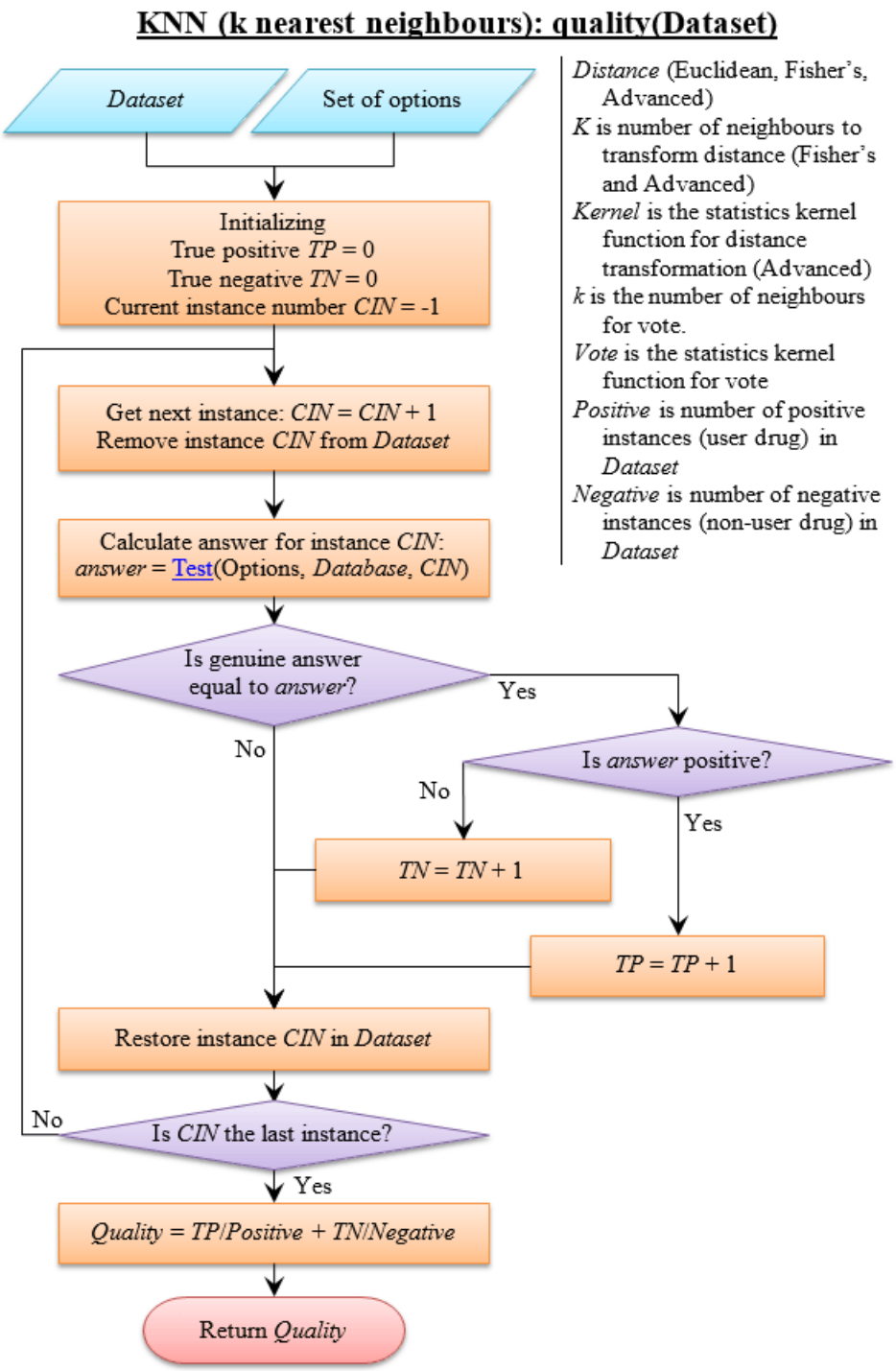

Figure A.8. kNN (k nearest neighbours): quality(Dataset).

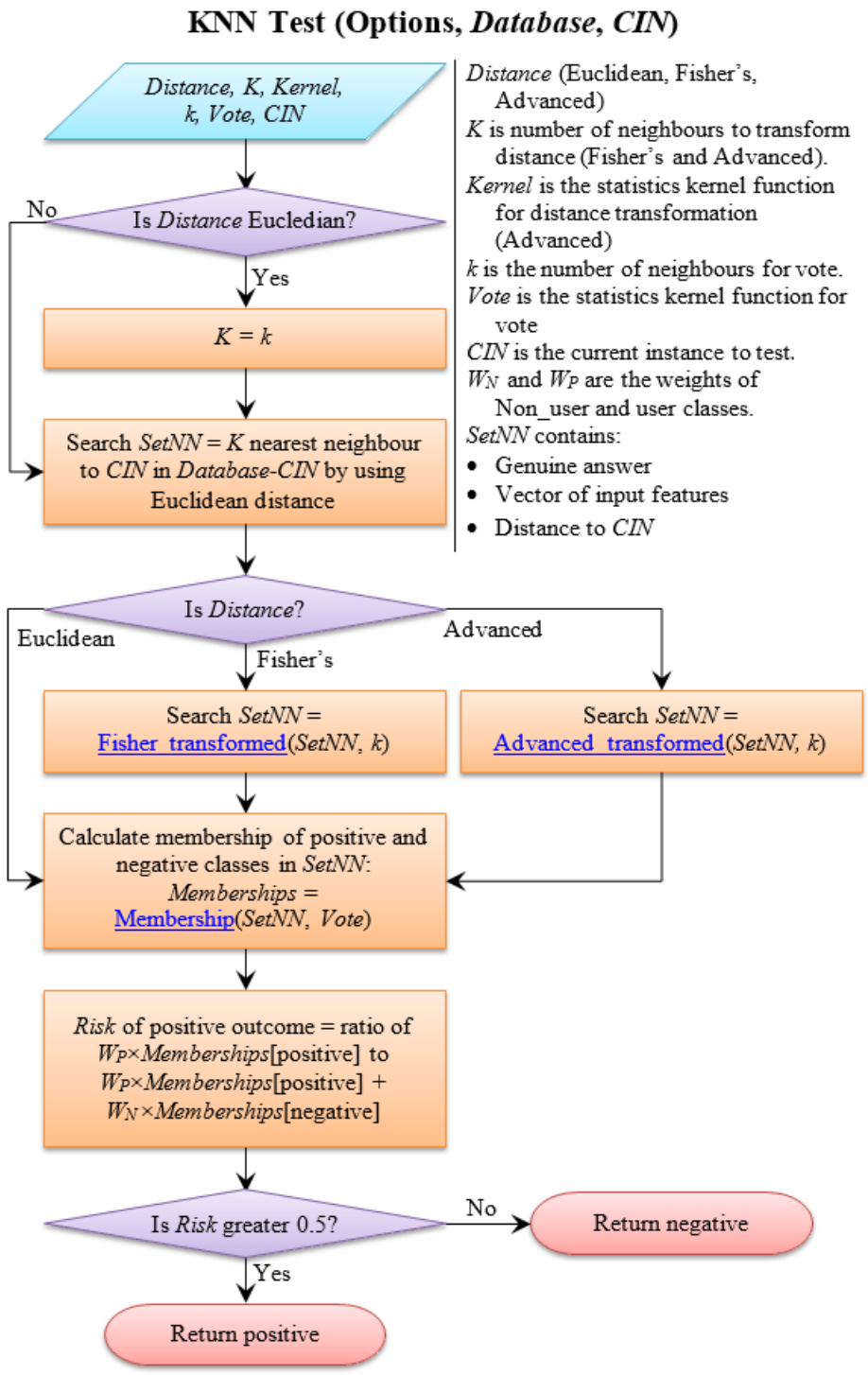

Figure A.9. KNN Test (Options, Database, CIN).

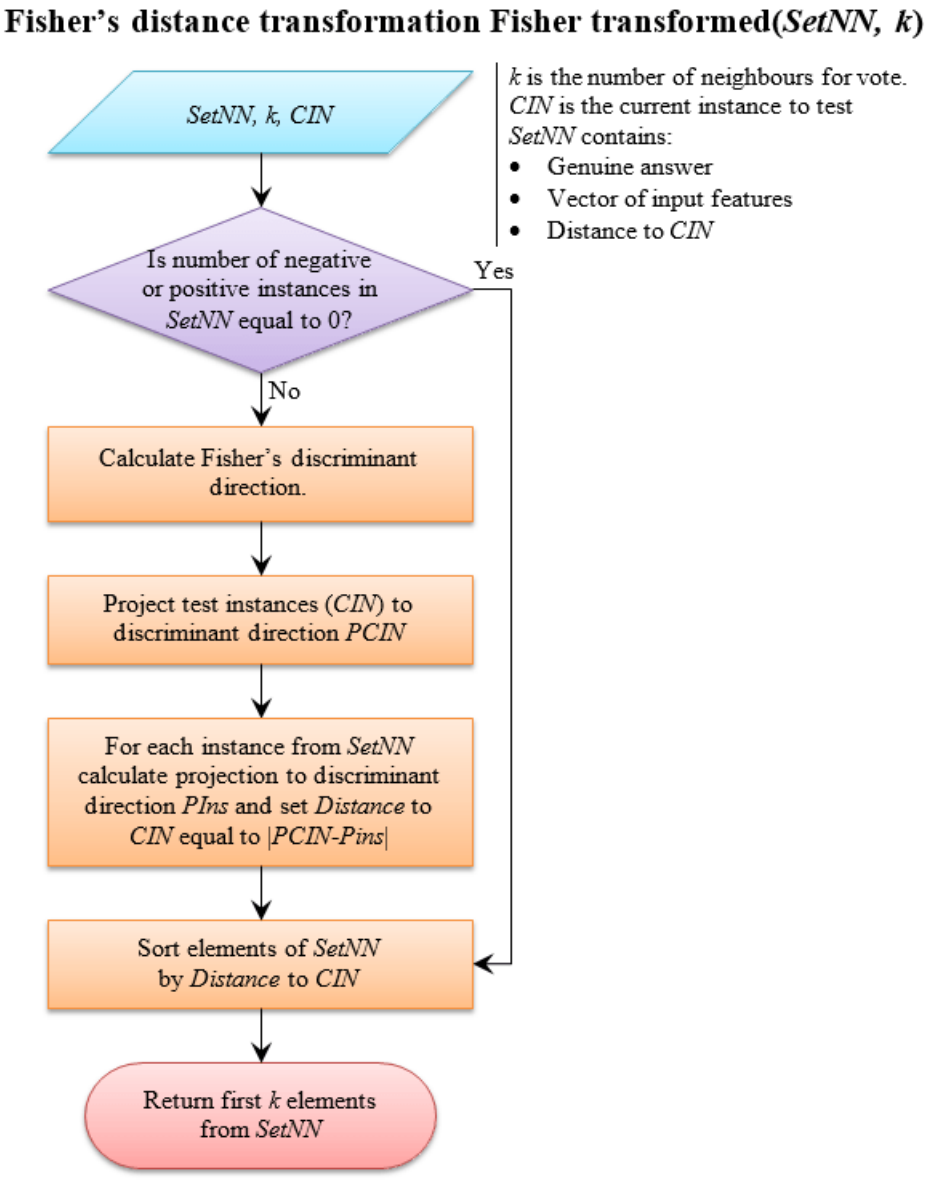

Figure A.10. Fisher's distance transformation Fisher transformed(SetNN, k).

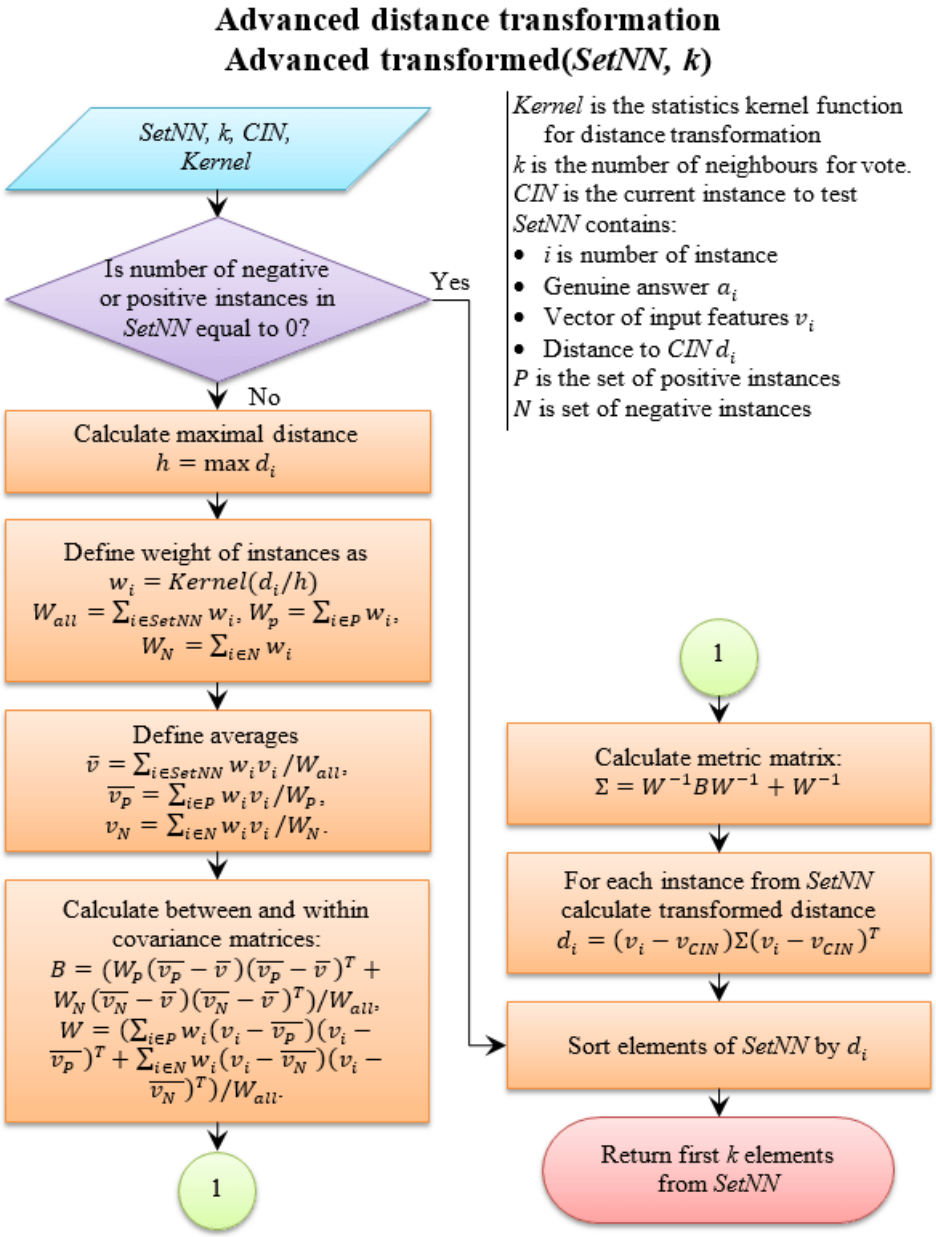

Figure A.11. Advanced distance transformation.

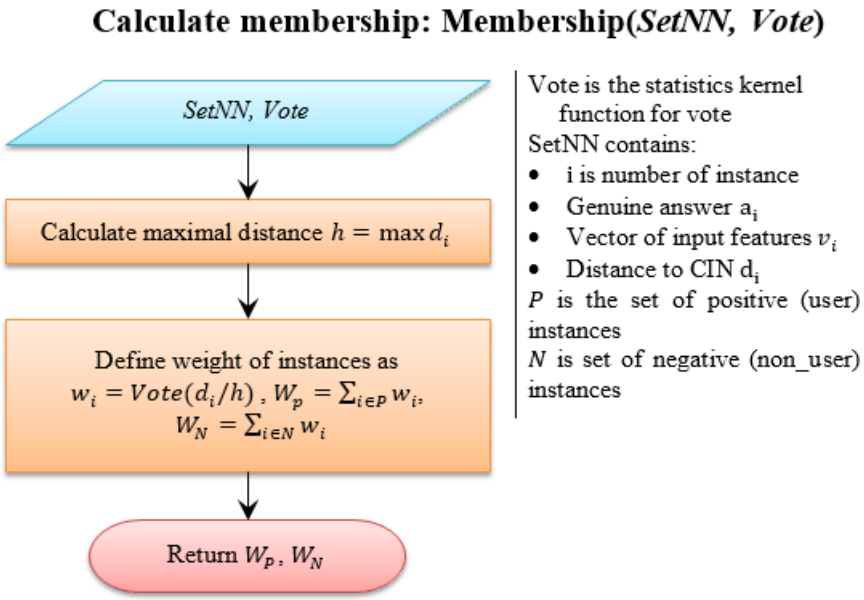

Figure A.12. Calculate membership: Membership(SetNN, Vote).

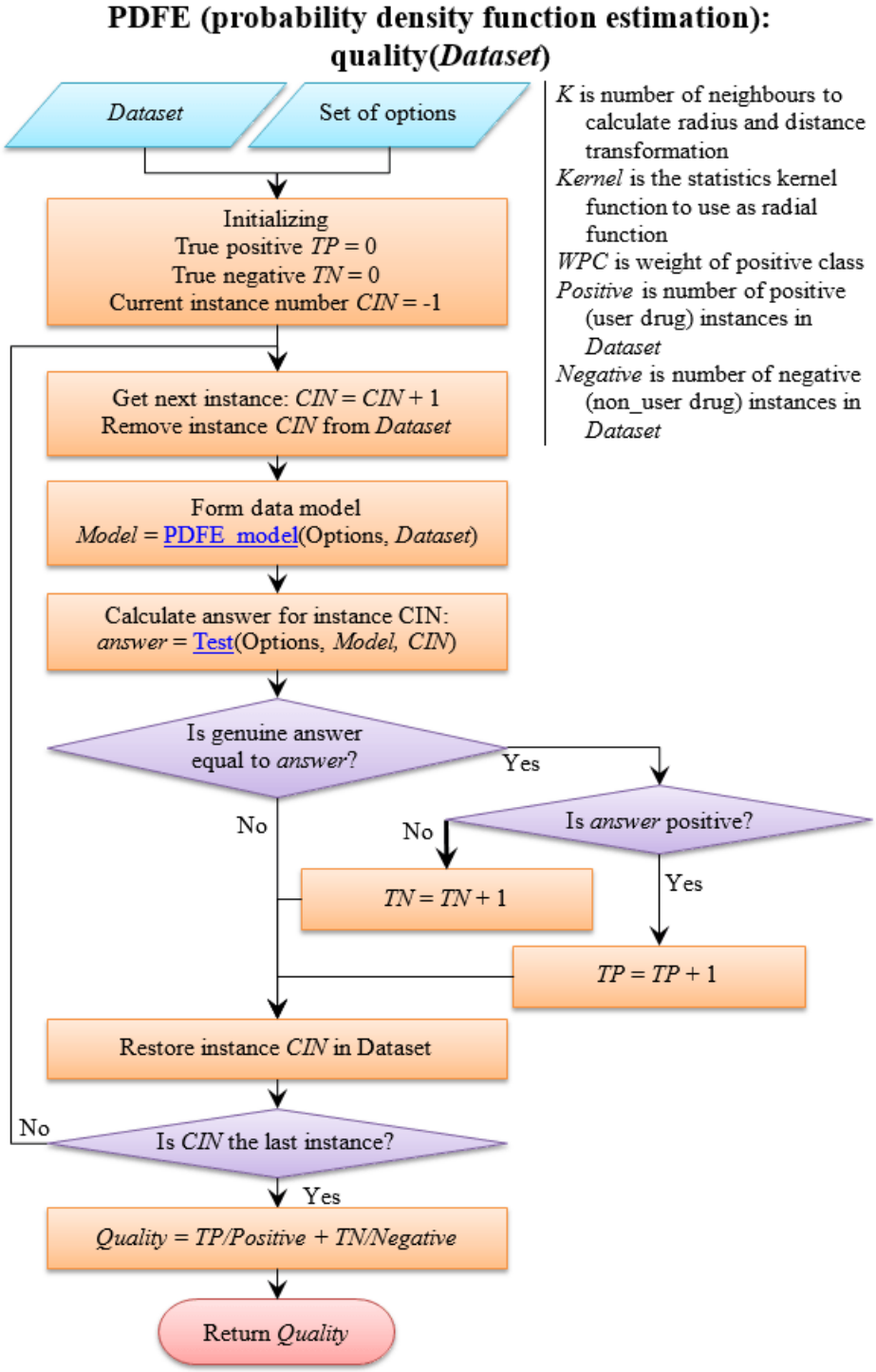

Figure A.13. PDFE (probability density function estimation): quality(Dataset).

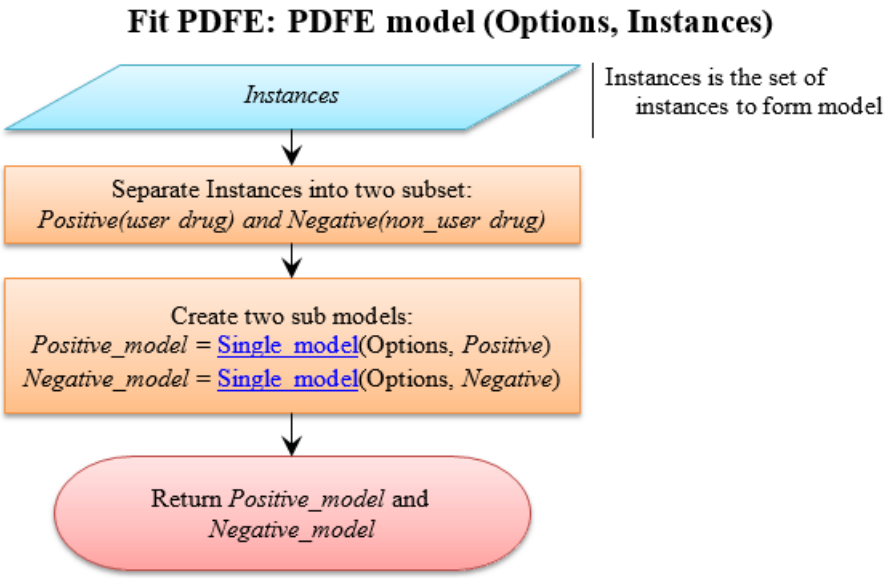

Figure A.14. Fit PDFE: PDFE model(Options, Instances).

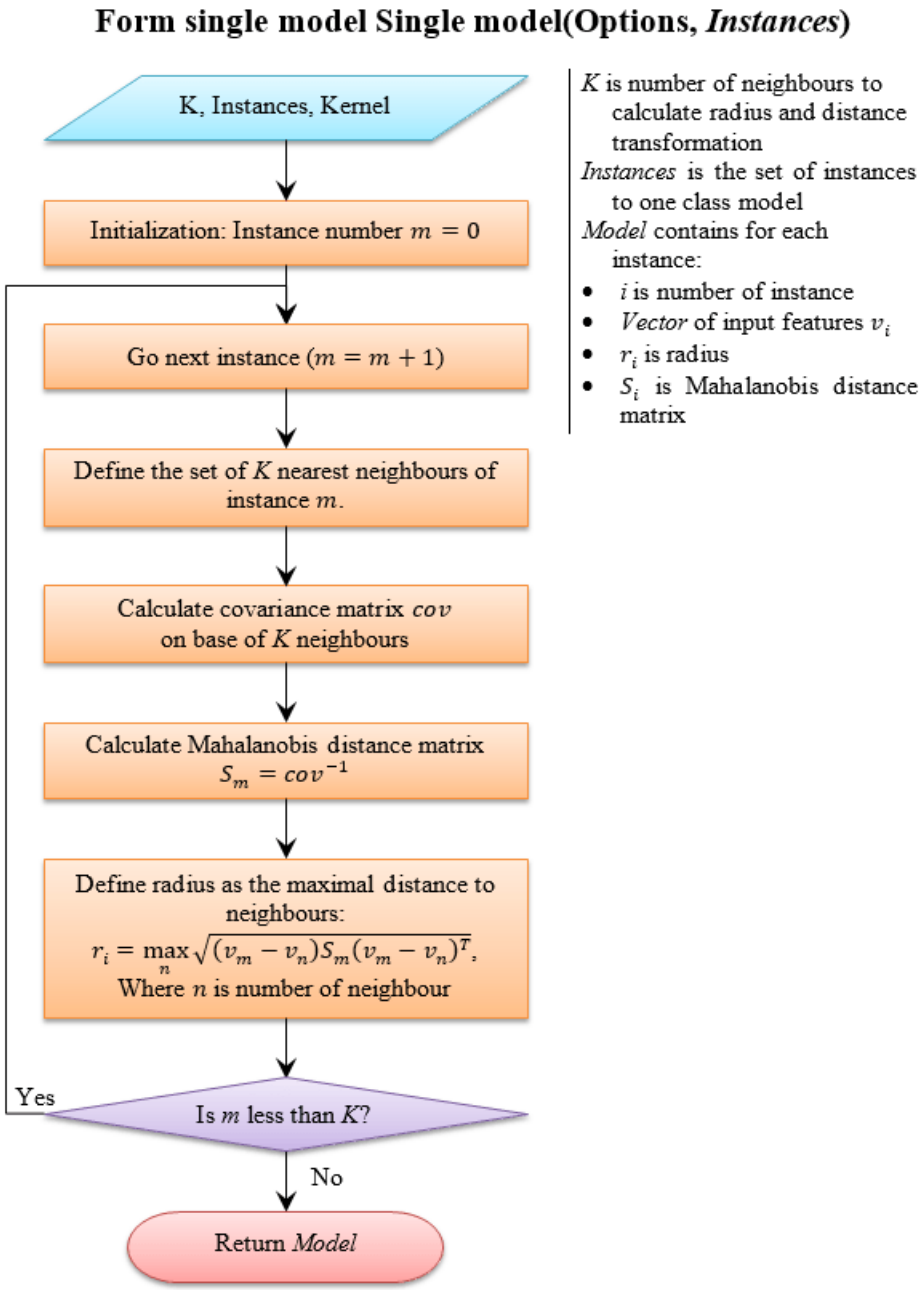

Figure A.15. Form single model Single model(Options, Instances).

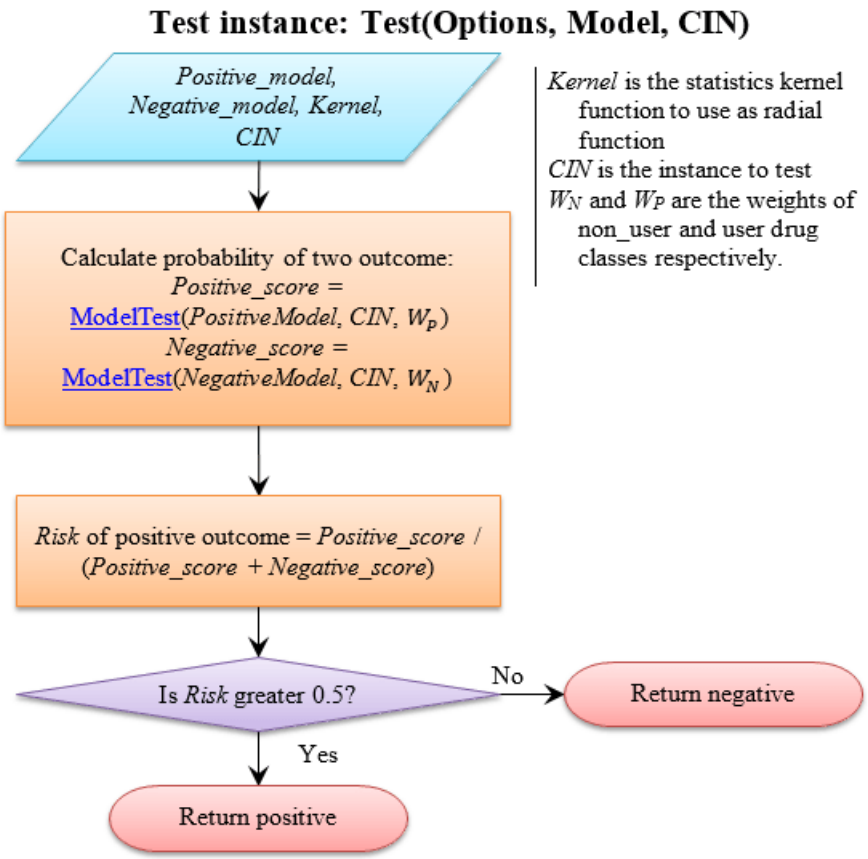

Figure A.16. Test instance: Test(Options, Model, CIN).

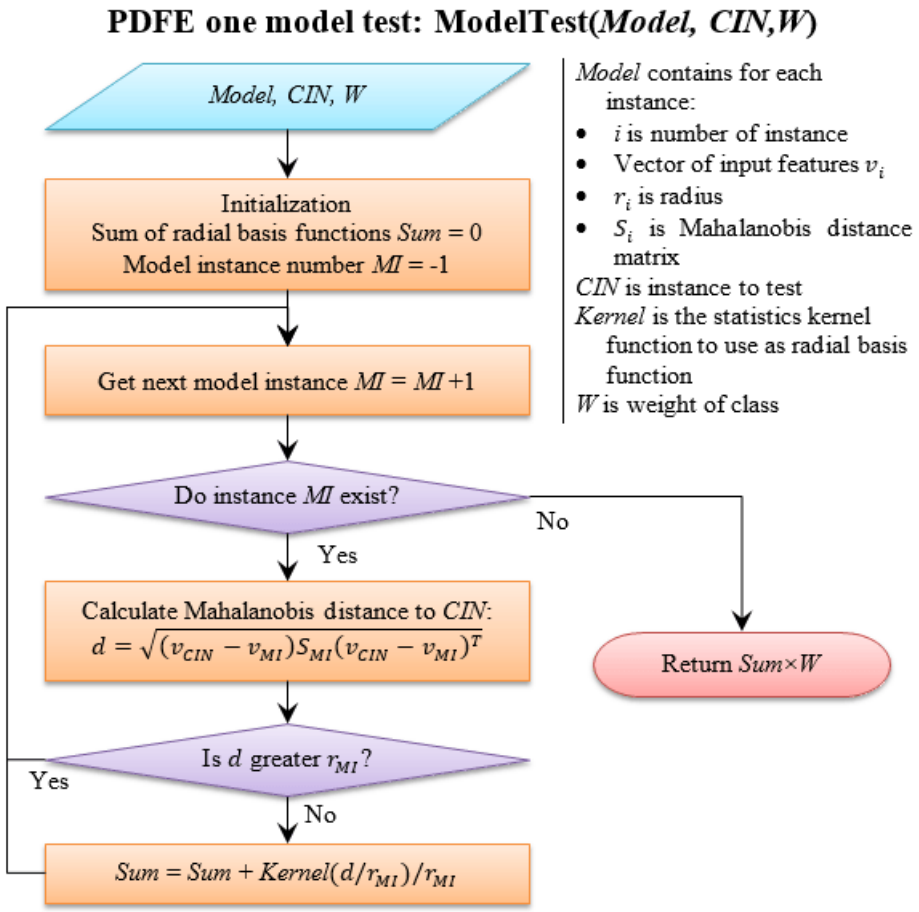

Figure A.17. PDFE one model test: ModelTest(Model, CIN,W).

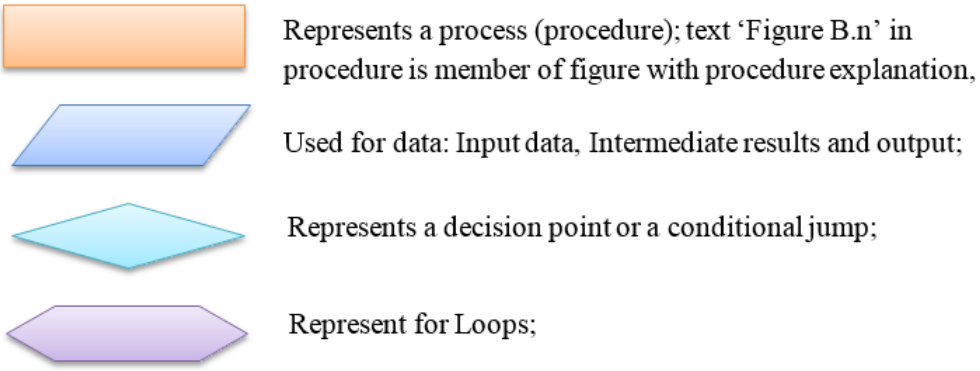

Figure B.1. Blocks of flowcharts.

APPENDIX B

Flowcharts of AASA for feature selection

We identify AASA by flowchart based on methods (ES, FFS, BFS) by using the classifiers. Flowchart title in form 'function name: Arg1, Arg2, ...' means usage of the same function with different arguments. Flowcharts use the following kinds of blocks:

Table B.1. List of abbreviations

Abbreviations	Meaning
FS	Feature Set
CF	Current Feature
CFS	Current Feature Set
Cl	Classifier
FSM	Feature Selection Method
ES	Exhaustive Search
FFS	Forward Feature Selection
BFS	Backward Feature Selection
FKA	First Kind Alternative
SKA	Second Kind Alternative
FSS	Feature Subset
OM	Optimal Model
OA	Optimal Accuracy
COM	Current Optimal Model
COA	Current Optimal Accuracy
OAM	Optimal Alternative Model
UFS	Union Feature Set
CM	Current Model
CA	Current Accuracy
MM	Minimal Model
MMA	Minimal Model Accuracy
OMM	Optimal Minimal Model
LOM	List Of Model
OMS	Optimal Minimal Set
LOMS	List of Optimal Minimal Set
EMCM	Ensemble Model Create Methods
CFS	Current Feature Set
RA	Required Accuracy

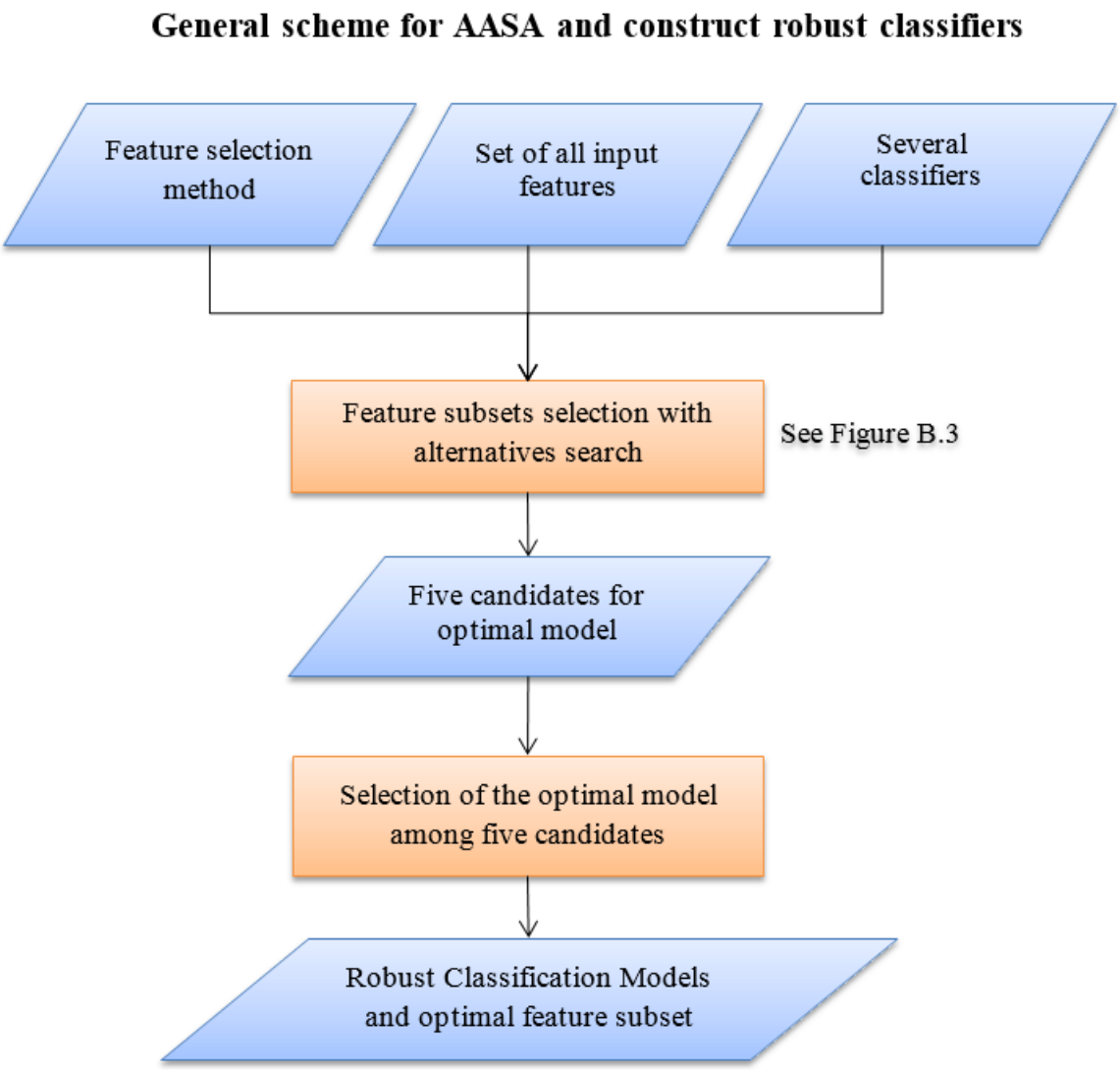

Figure 1. General scheme for AASA.

Figure B.2. General scheme for AASA and construct robust classifiers.

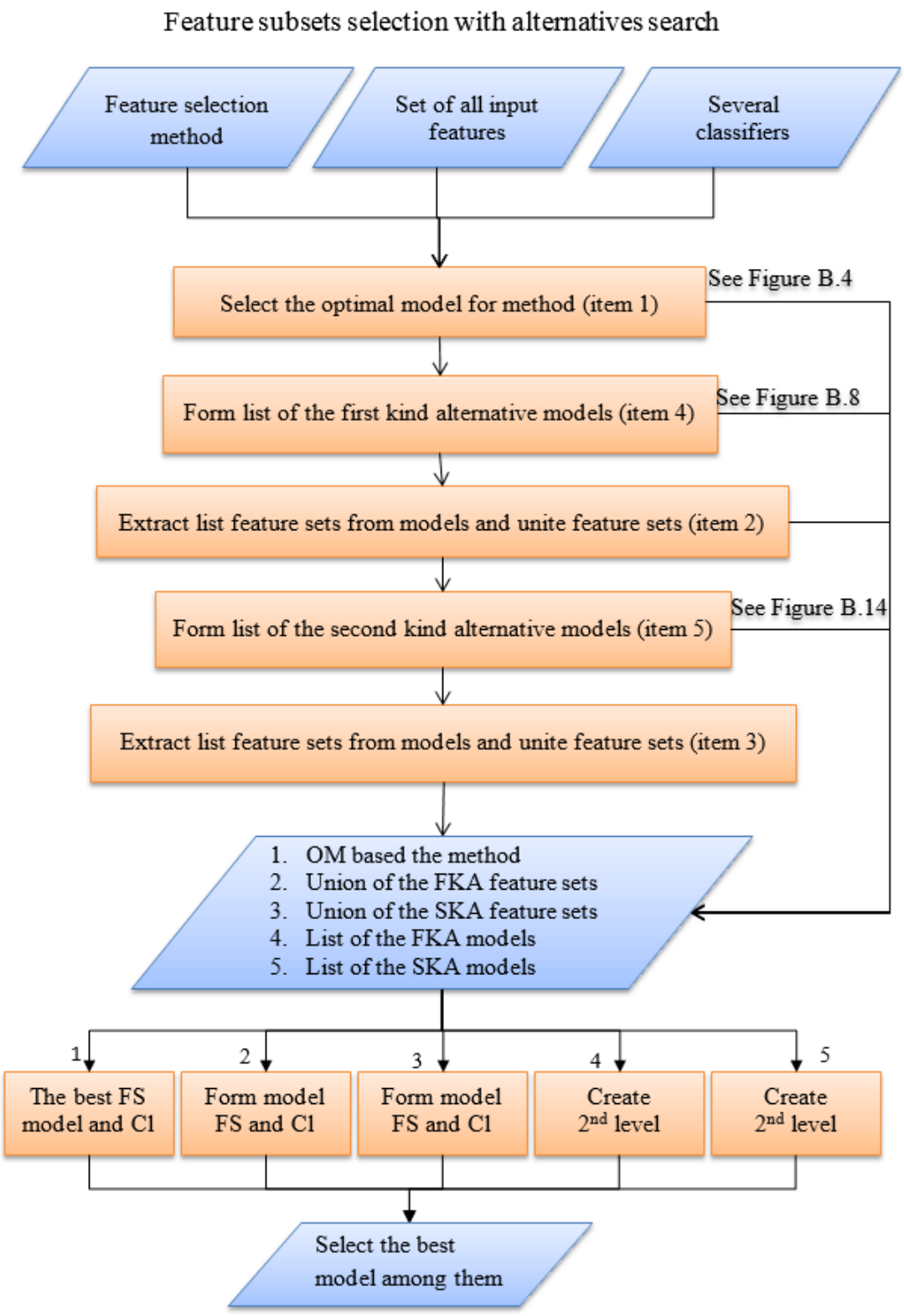

Figure B.3. Feature subsets selection with AAS.

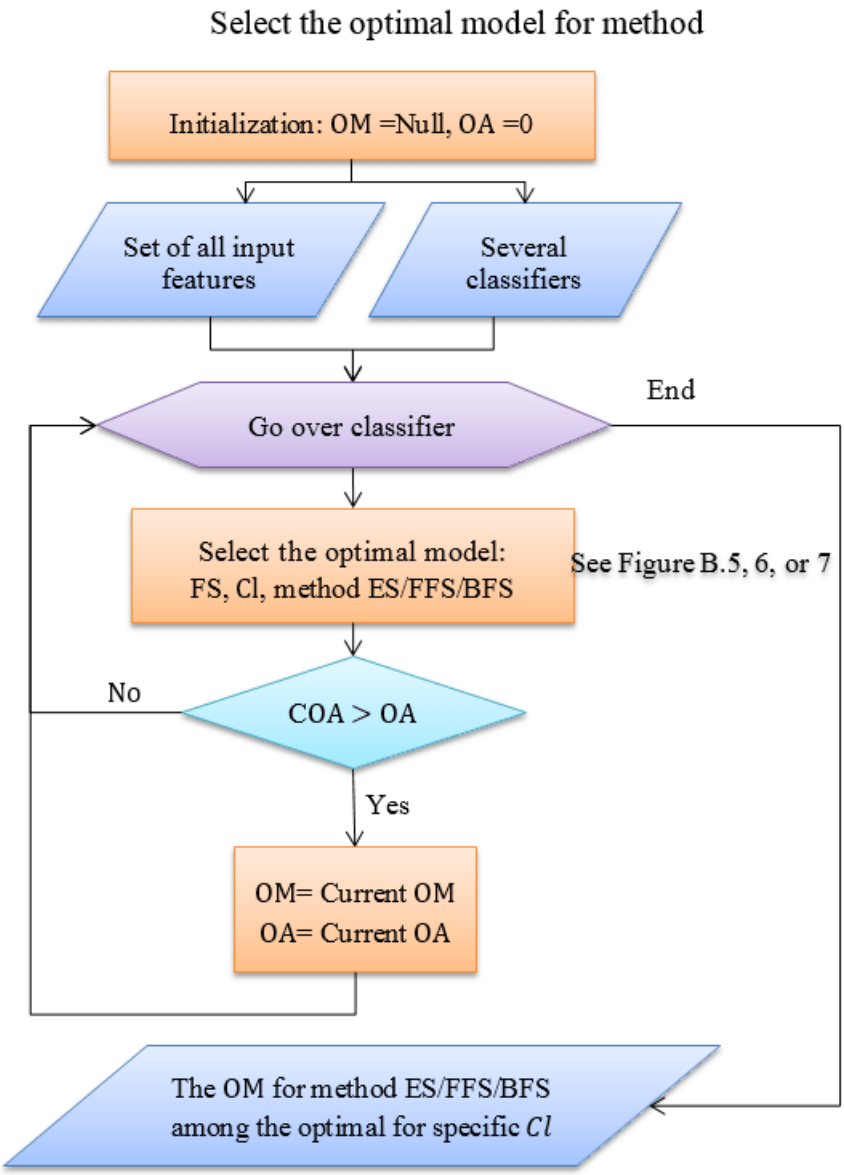

Figure B.4. Select the optimal model.

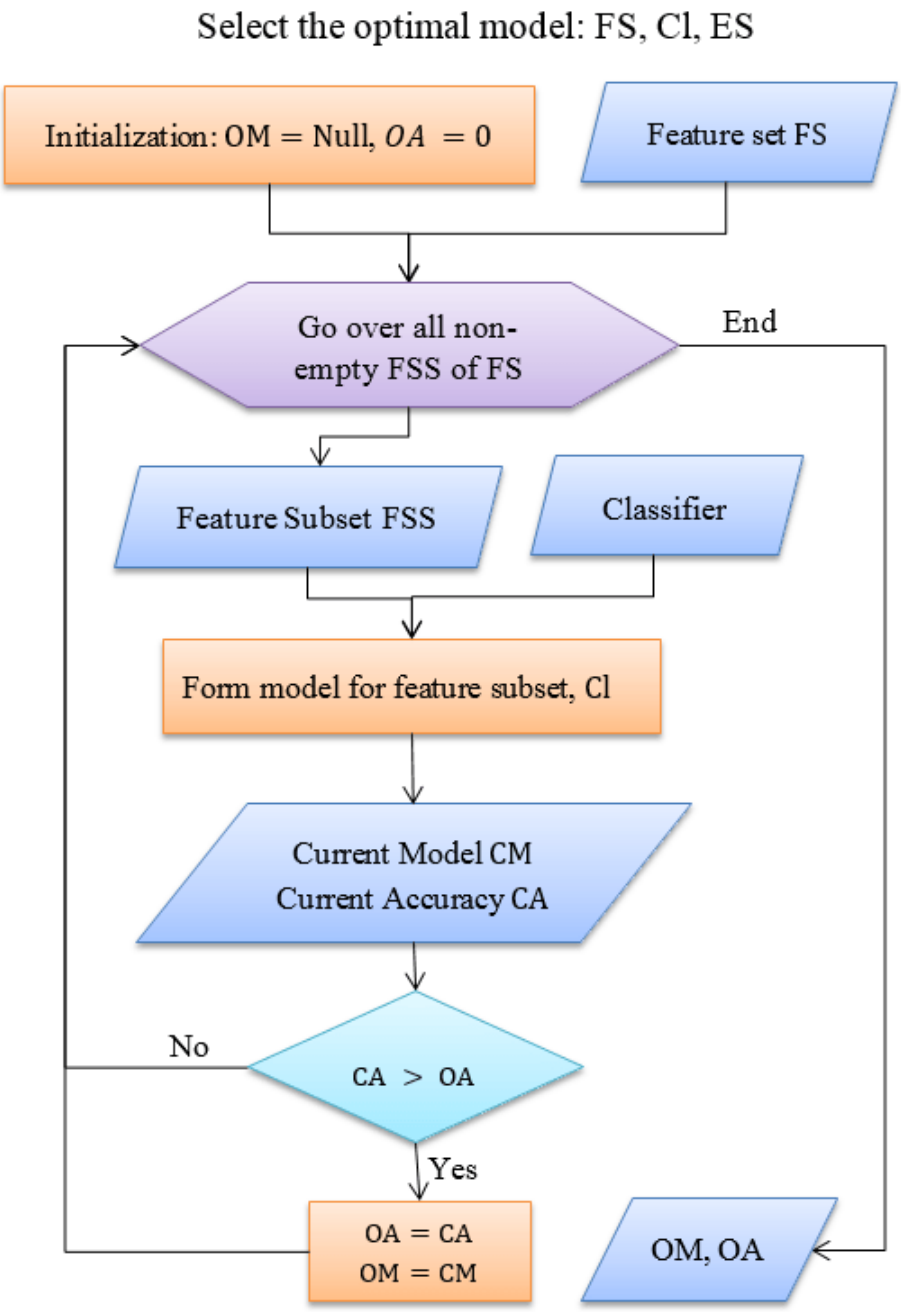

Figure B.5. Select the optimal model: FS, Cl, ES.

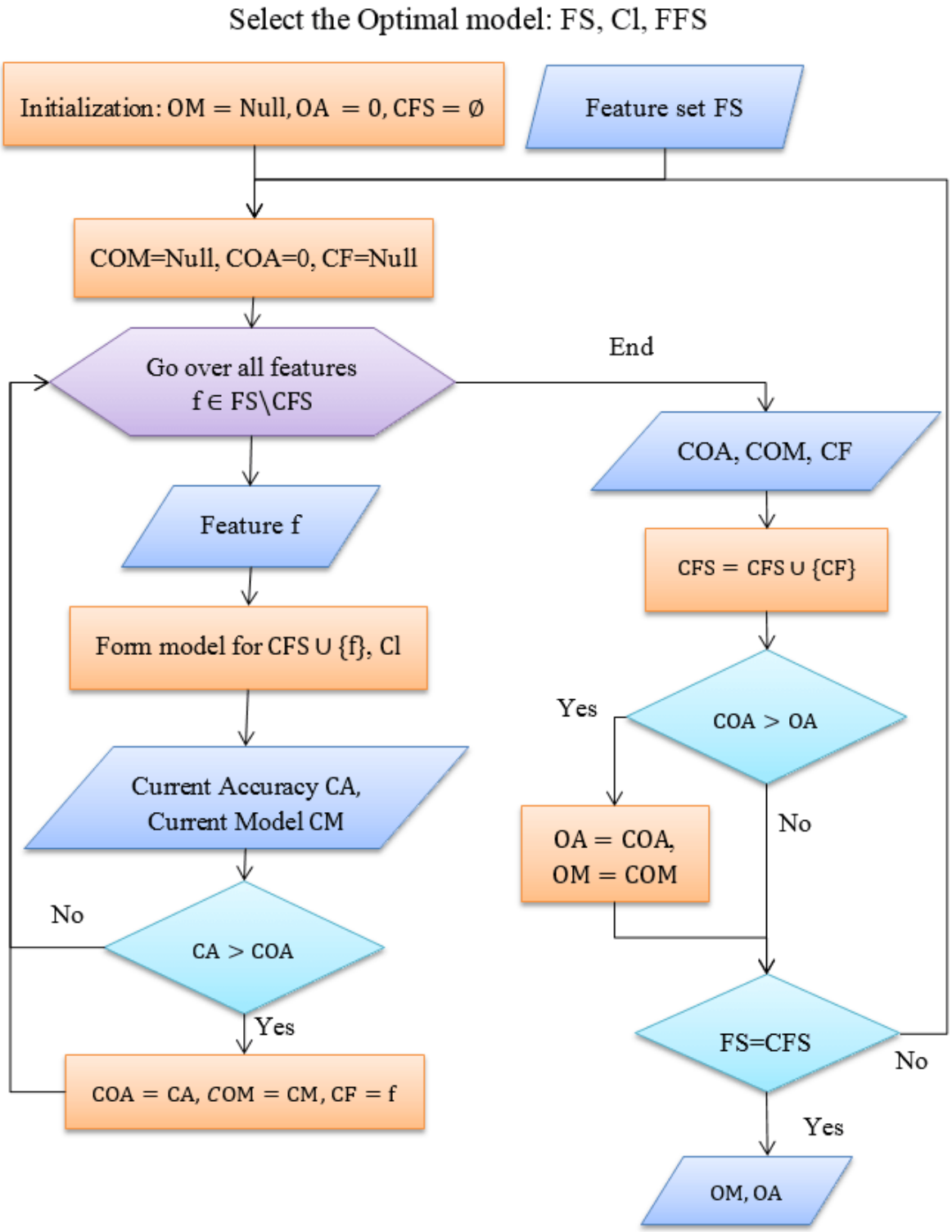

Figure B.6. Select the Optimal model: FS, Cl, FFS.

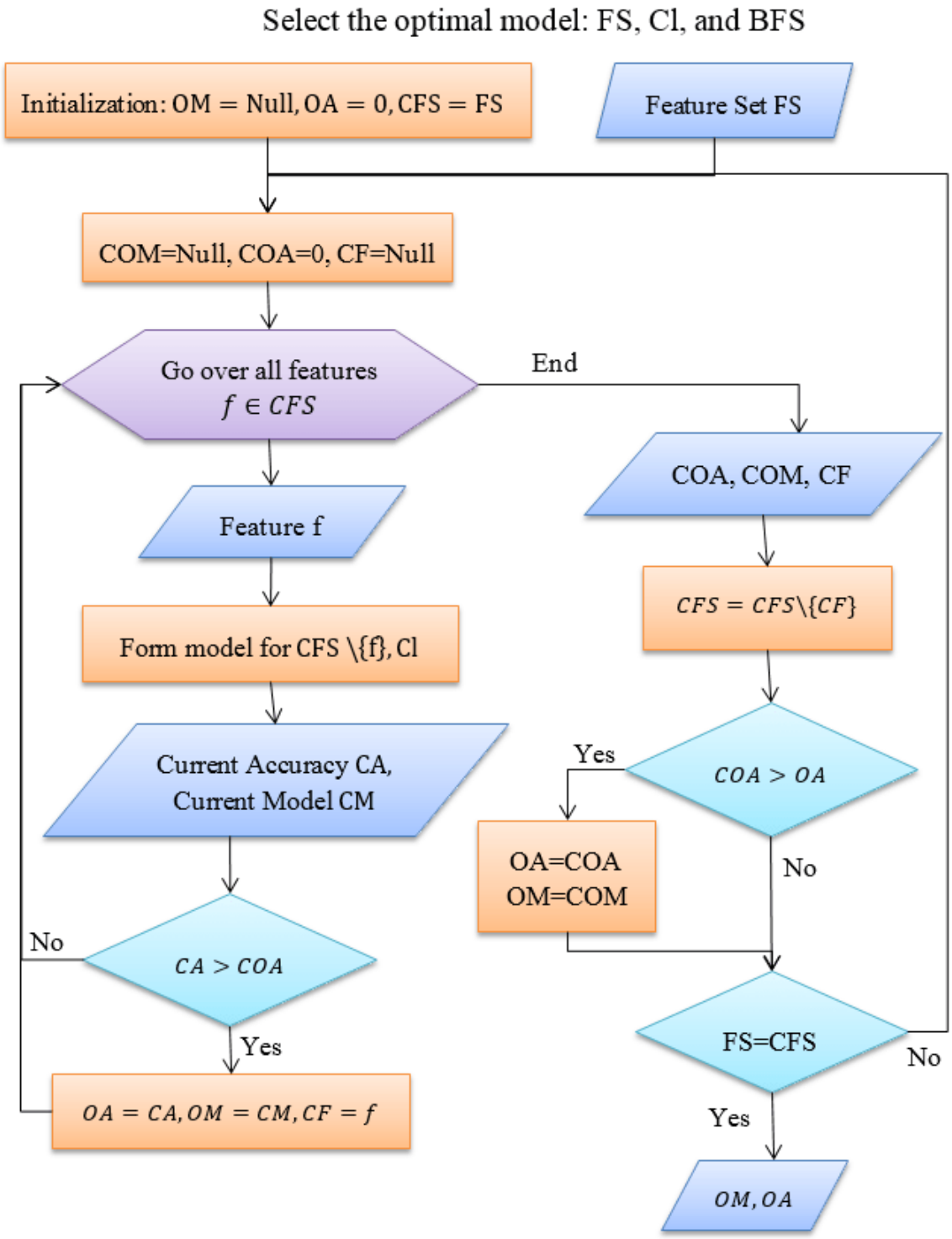

Figure B.7. Select the optimal model: FS, Cl, BFS.

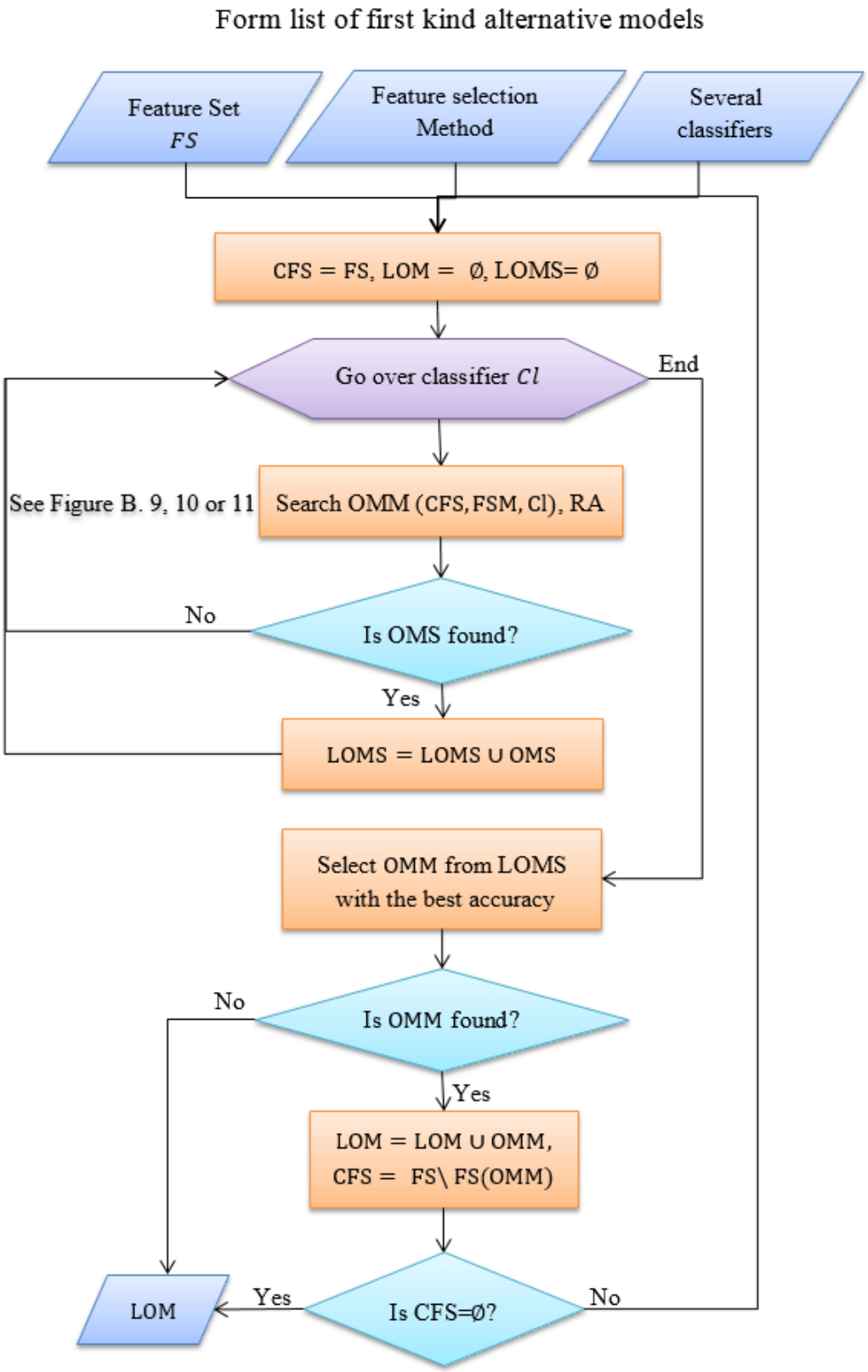

Figure B.8. Form list of first kind AASA models.

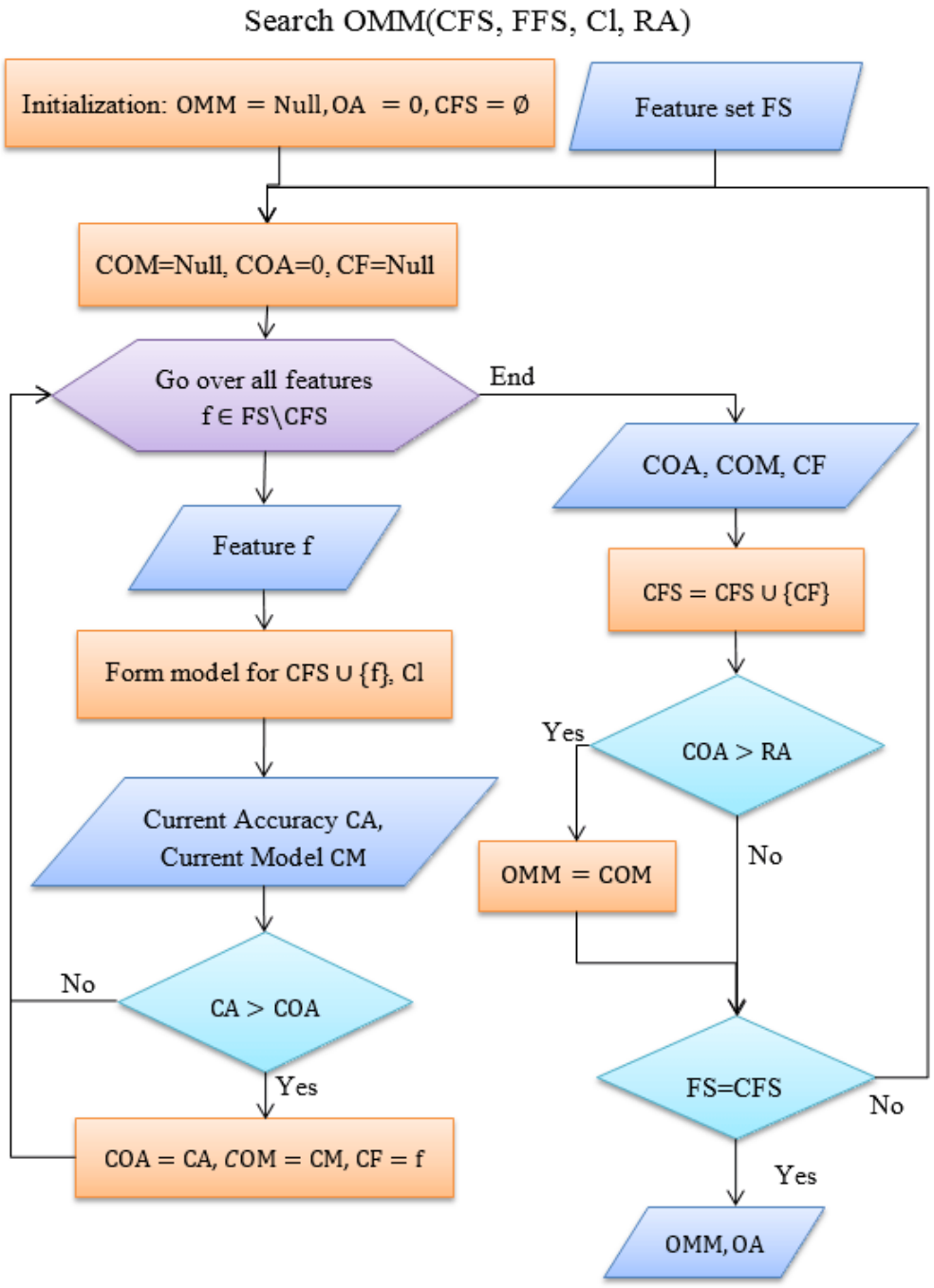

Figure B.9. Search OMM (CFS, FFS, and Cl, RA).

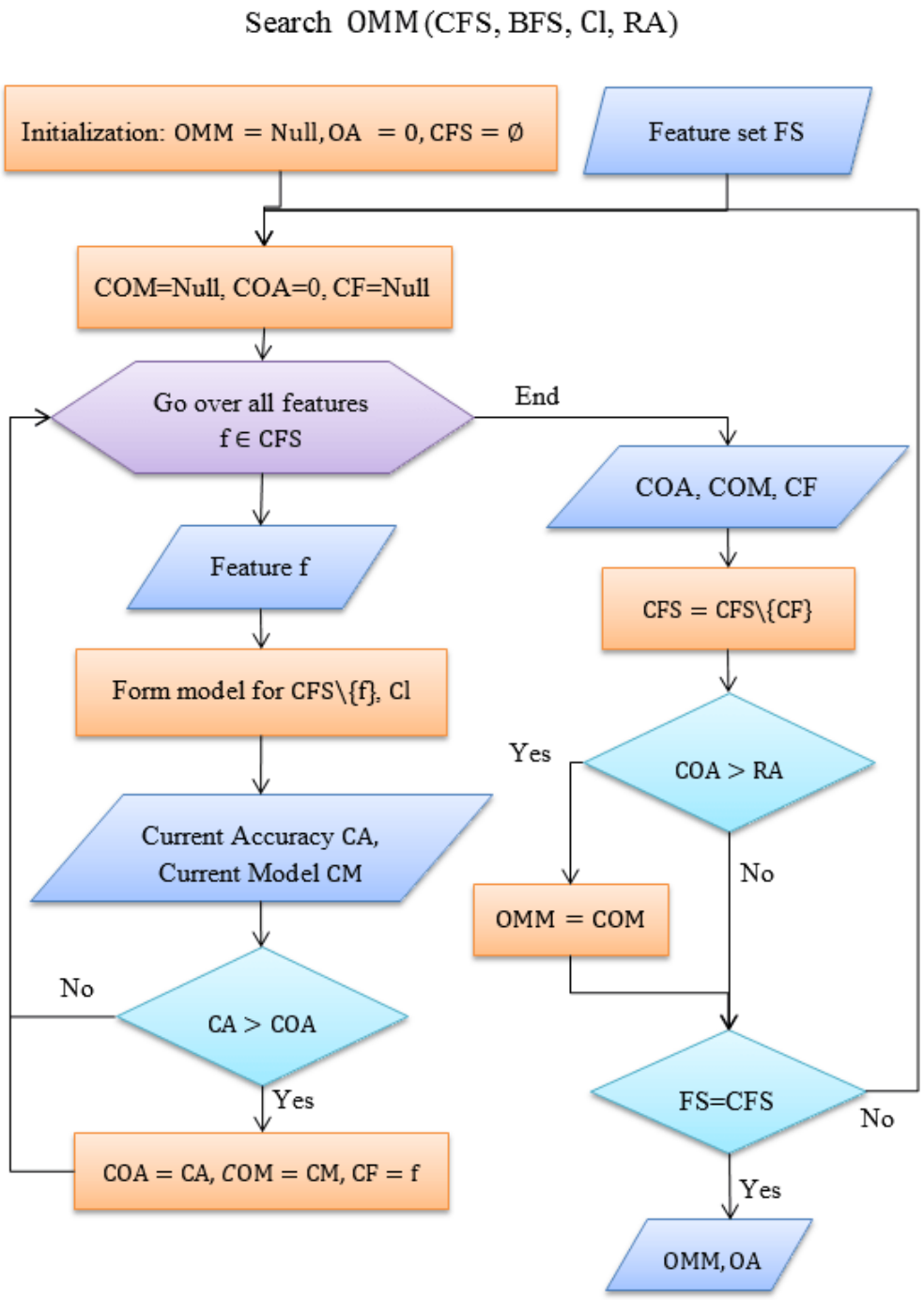

Figure B.10. Search OMM (CFS, BFS, Cl, RA).

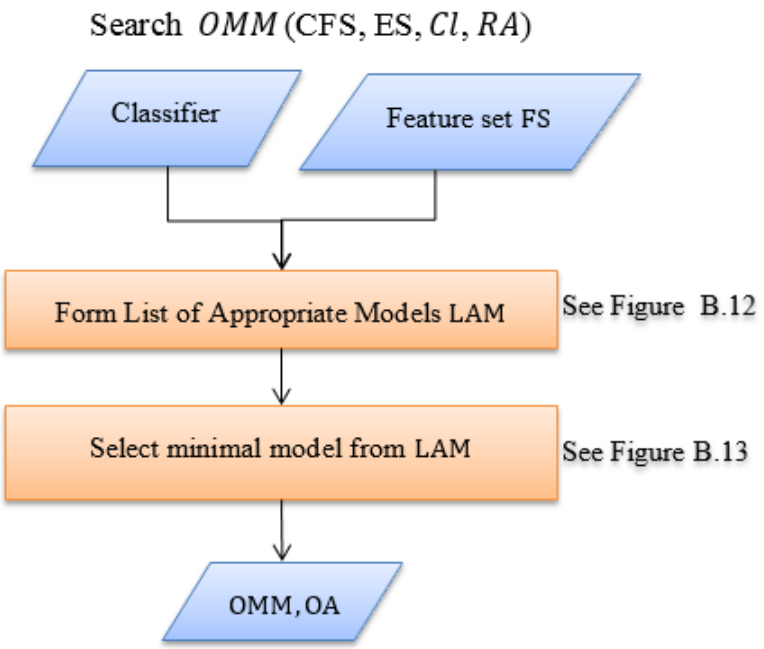

Figure B.11. Search OMM (CFS, ES, Cl, RA).

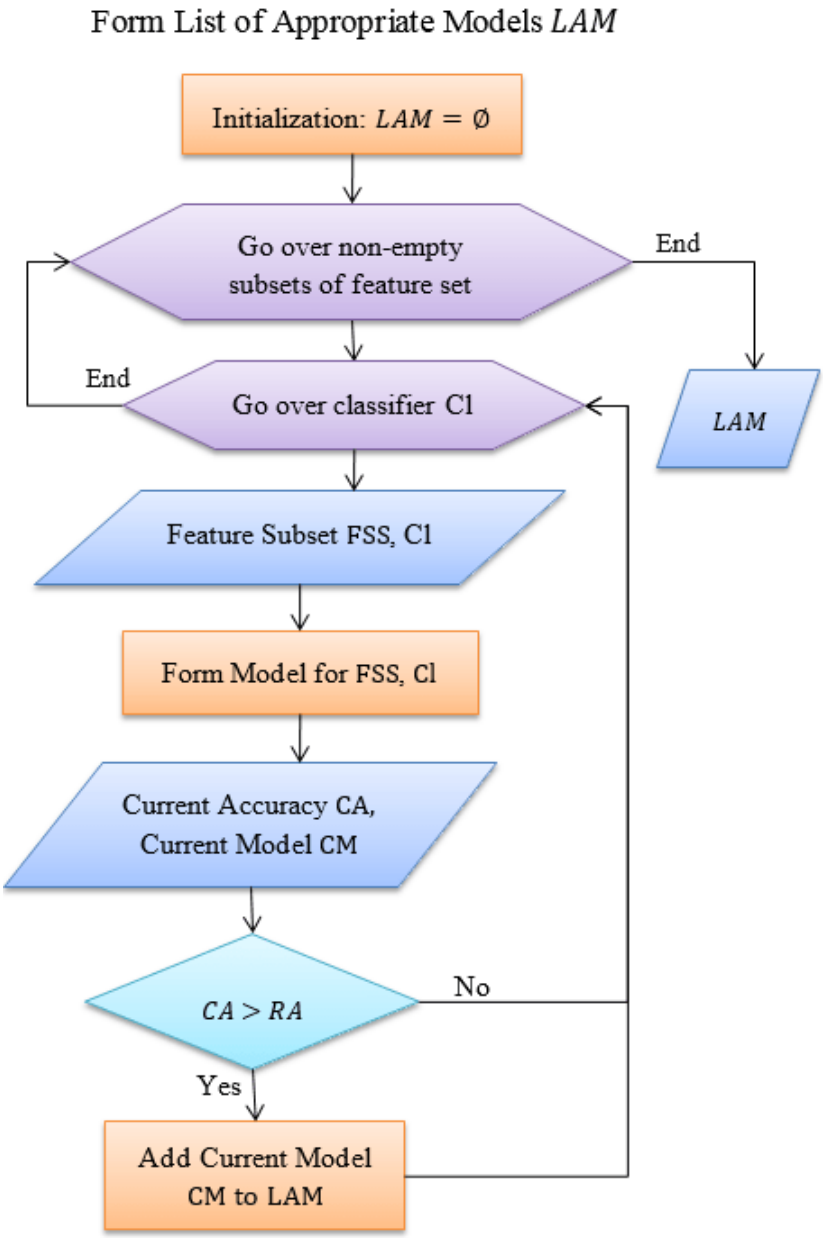

Figure B.12. Form List of Appropriate Models (LAM).

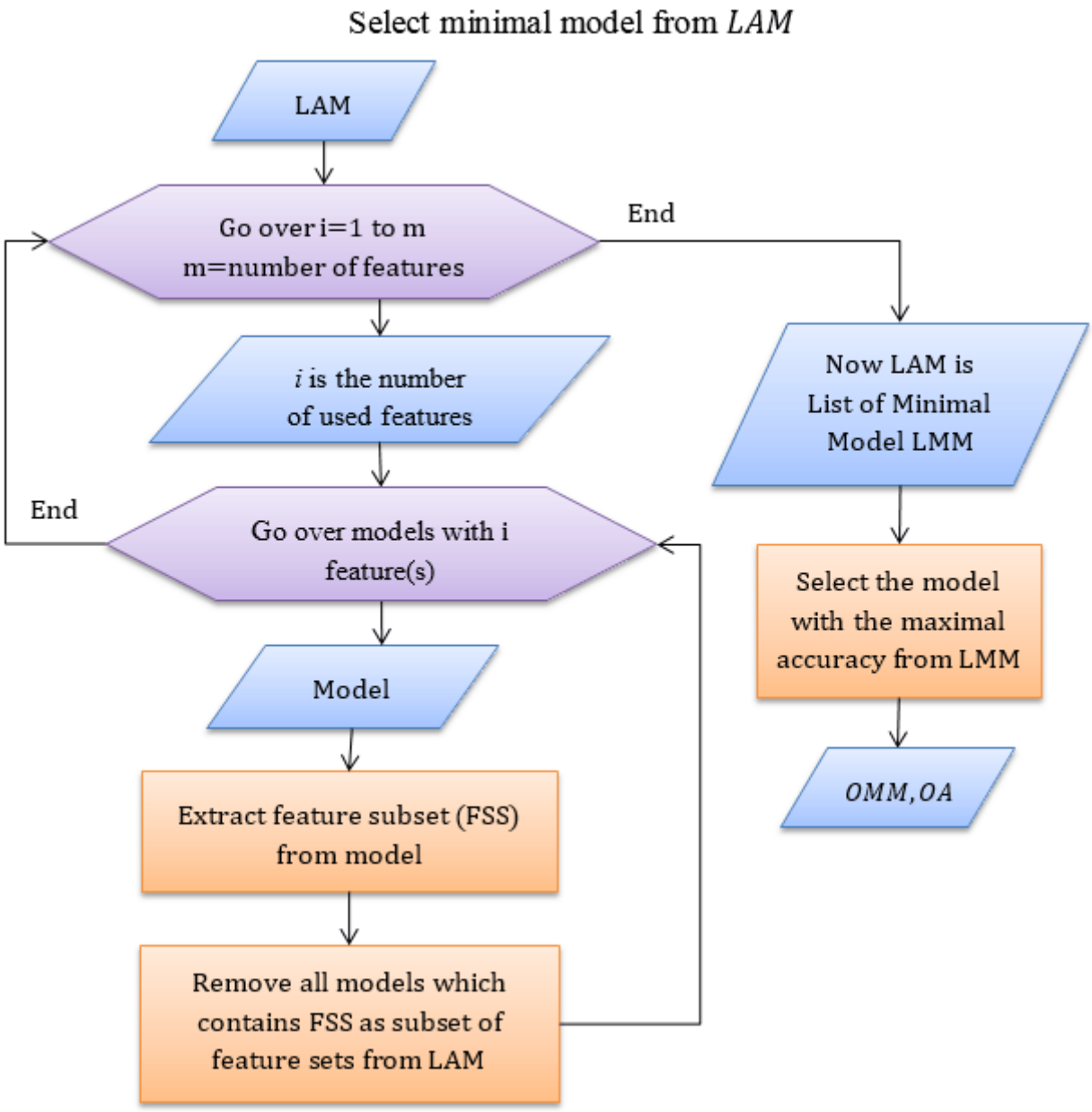

Figure B.13. Select minimal model from LAM.

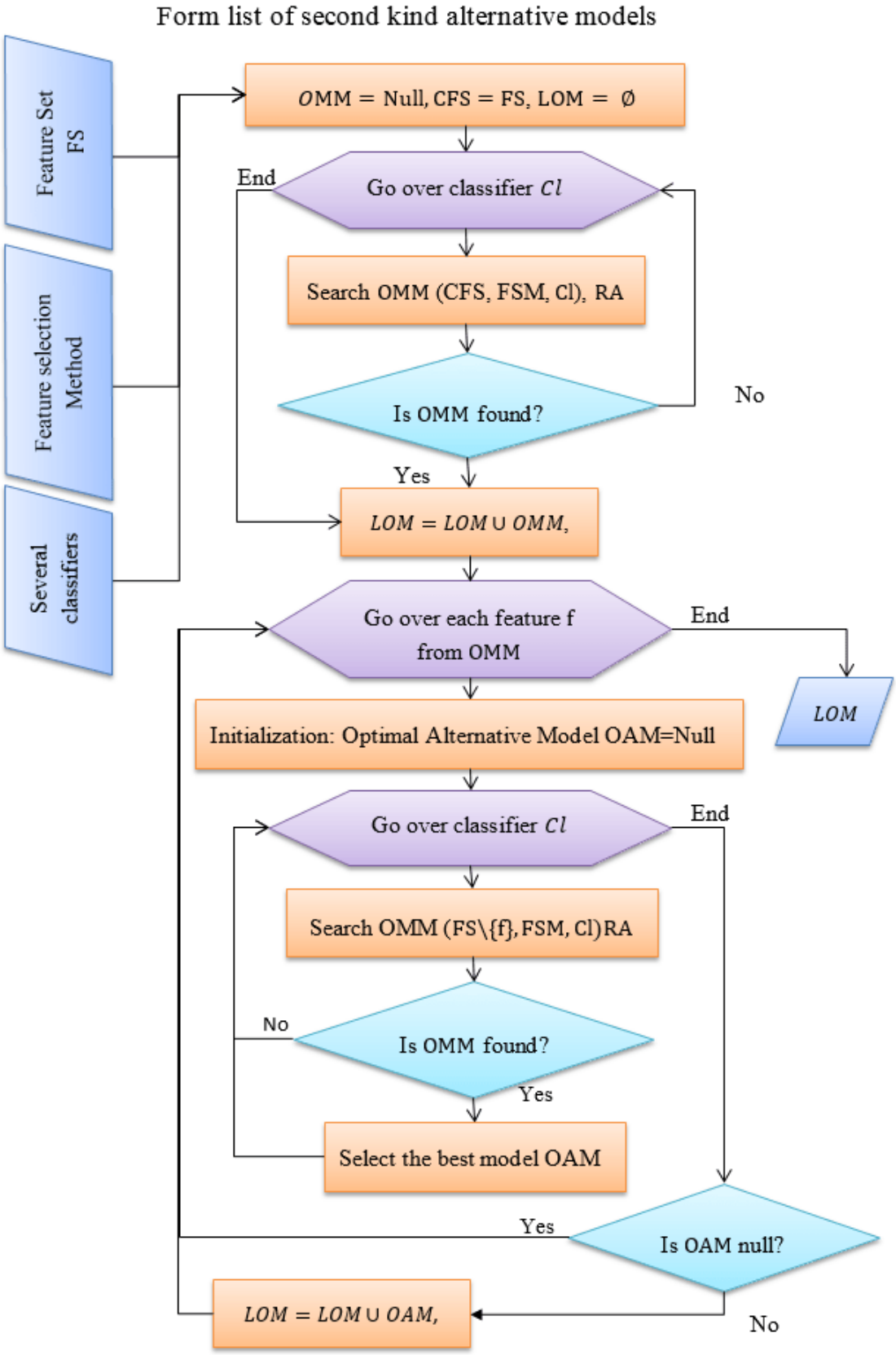

Figure B.14. Form list of second kind AASA models.

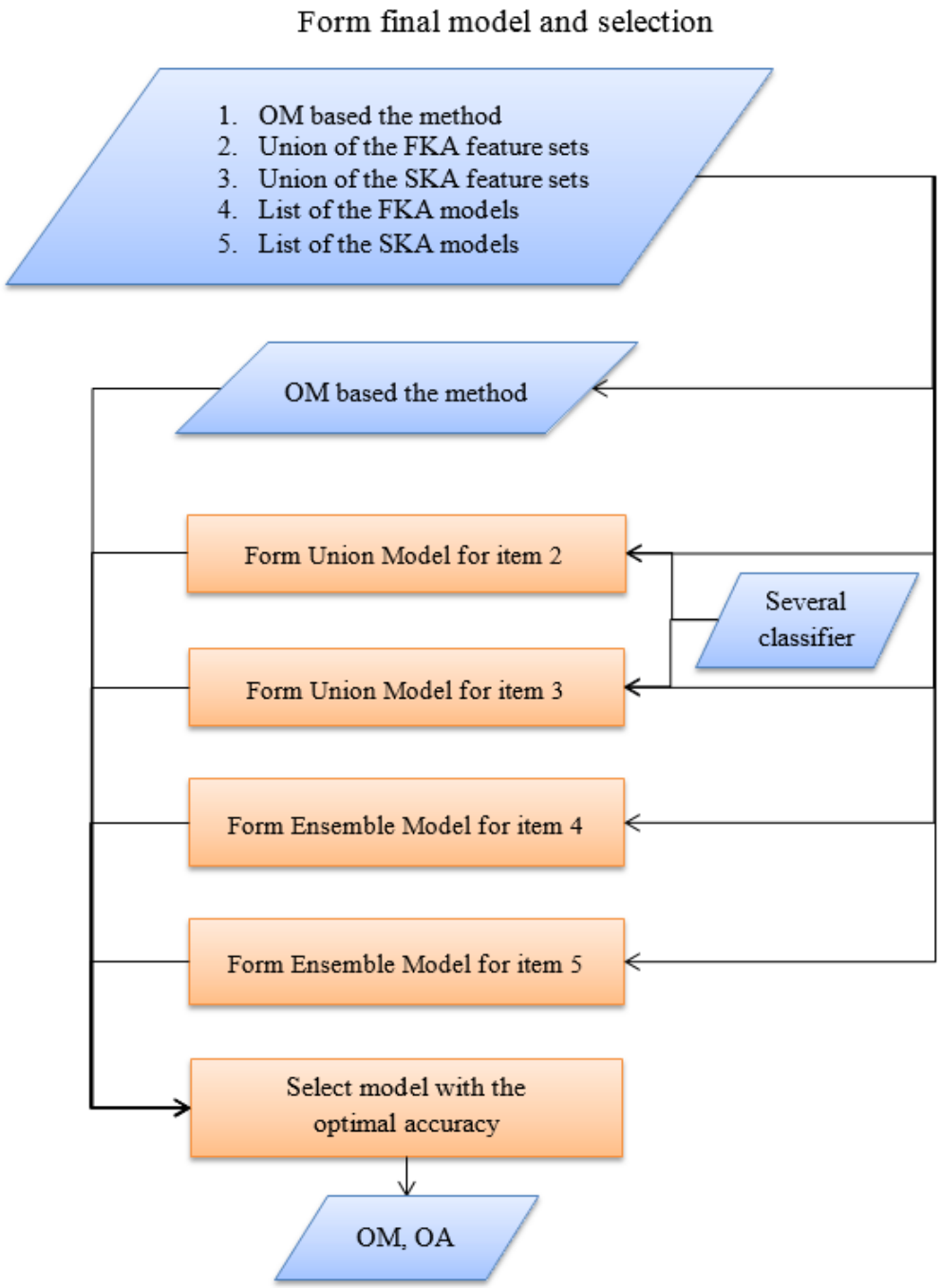

Figure B.15. Selection of the optimal model among five candidates.

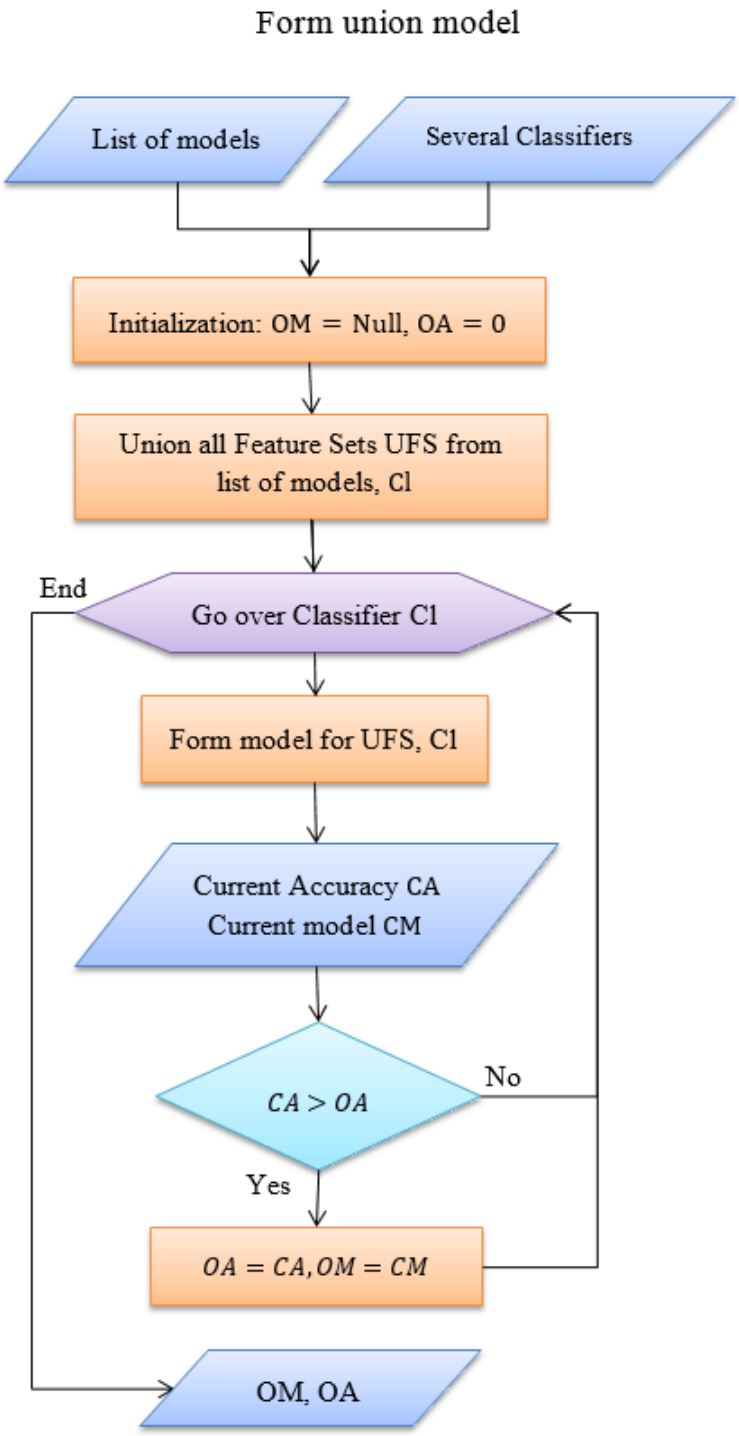

Figure B.16. Form union model.

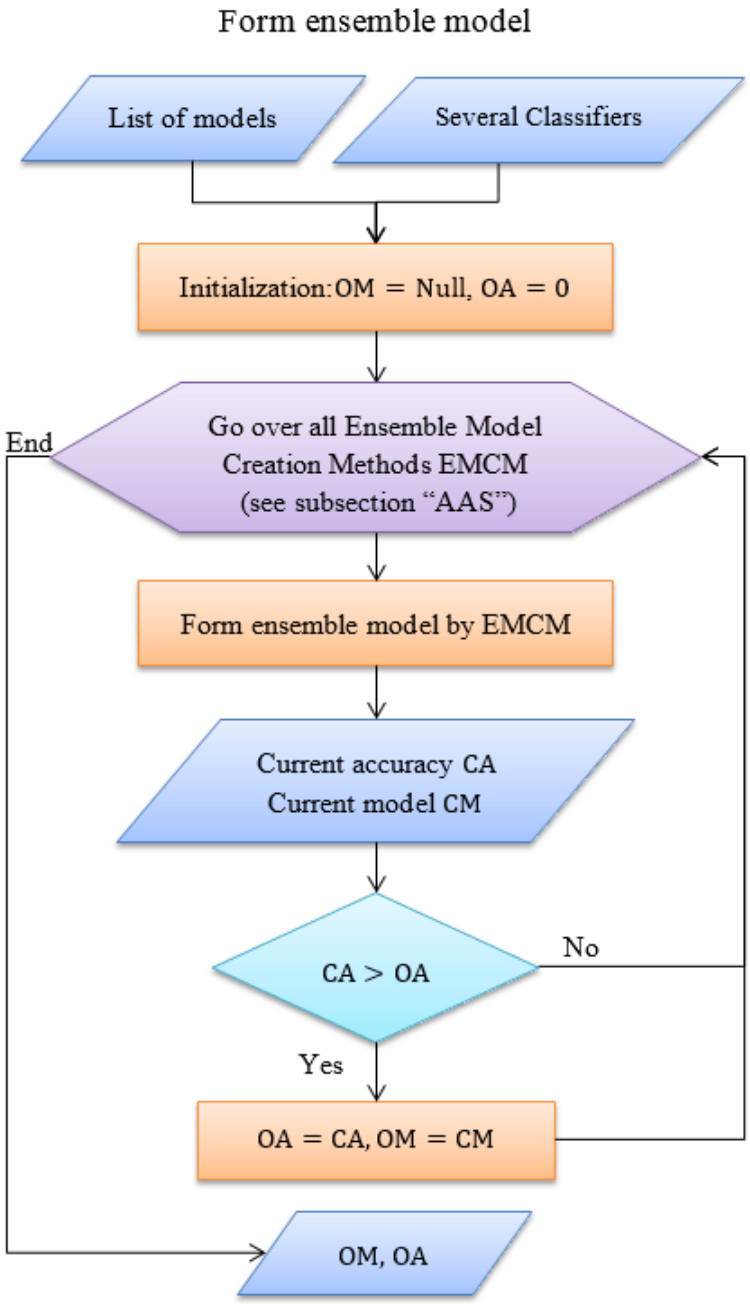

Figure B.17. Form ensemble model.

APPENDIX C

Main tables

C.1 Psychological profiles of drug users and non-users

Mean for groups of users and non-users. In this Appendix, we present mean T-$score_{sample}$ (3.1.2) for groups of users and non-users for decade, year, month, and week-based user definitions respectively. Column p-value assesses the significance of differences of mean scores for groups of users and non-users: it is the probability of observing by chance the same or greater differences for mean scores if both groups have the same mean. Rows '#' contain number of users and non-users for the drugs. These tables include all information about five factor personality profiles for various definitions of users of 18 drugs, and four groups of drugs:

- The heroin pleiad: crack, cocaine, methadone, and heroin;
- The ecstasy pleiad: amphetamines, cannabis, cocaine, ketamine, LSD, magic mushrooms, legal highs, and ecstasy;
- The benzodiazepines pleiad includes methadone, amphetamines, cocaine, and benzodiazepines.
- The union of these three pleiades and VSA: amphetamines, amyl nitrite, benzodiazepines, cannabis, cocaine, crack, ecstasy, heroin, ketamine, le-

gal highs, LSD, methadone, magic mushrooms (MMushrooms), and VSA, which we call for short 'illicit drugs' (with some abuse of language).

Table C.1. Mean T-$score_{sample}$ (MT) and 95% CI for it for groups of users and non-users with decade-based definition of users

Factor	Users		Non-users		p-value
	MT	95% CI	MT	95% CI	
Alcohol					
#	1817		68		
N	50.07	49.61, 50.53	48.12	45.70, 50.54	0.119
E	50.00	49.54, 50.46	50.00	47.63, 52.36	0.997
O	50.04	49.58, 50.50	48.86	46.51, 51.22	0.331
A	49.91	49.45, 50.37	52.49	50.24, 54.74	0.028
C	49.88	49.42, 50.34	53.28	51.01, 55.54	0.004
Imp	50.14	49.68, 50.60	46.23	44.09, 48.37	0.001
SS	50.19	49.74, 50.65	44.81	42.57, 47.06	< 0.001
Amphetamine					
#	679		1206		
N	51.64	50.88, 52.39	49.08	48.52, 49.64	< 0.001
E	49.65	48.84, 50.47	50.20	49.66, 50.73	0.274
O	53.05	52.34, 53.76	48.28	47.72, 48.84	< 0.001
A	48.37	47.58, 49.16	50.92	50.37, 51.46	< 0.001
C	46.97	46.21, 47.73	51.71	51.17, 52.25	< 0.001
Imp	53.49	52.76, 54.22	48.04	47.49, 48.58	< 0.001
SS	54.76	54.12, 55.40	47.32	46.77, 47.87	< 0.001
Amyl nitrite					
#	370		1515		
N	50.72	49.72, 51.73	49.82	49.32, 50.33	0.118
E	50.89	49.87, 51.91	49.78	49.28, 50.29	0.055
O	51.45	50.48, 52.42	49.65	49.14, 50.15	0.001

Continued on the next page

Table C.1. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
A	48.65	47.62, 49.69	50.33	49.83, 50.83	0.004
C	48.01	47.02, 48.99	50.49	49.98, 50.99	< 0.001
Imp	52.20	51.23, 53.18	49.46	48.96, 49.97	< 0.001
SS	53.89	53.02, 54.77	49.05	48.54, 49.56	< 0.001
Benzodiazepines					
#	769		1116		
N	52.77	52.06, 53.48	48.09	47.53, 48.65	< 0.001
E	49.02	48.27, 49.78	50.67	50.12, 51.23	0.001
O	52.64	51.96, 53.33	48.18	47.60, 48.76	< 0.001
A	48.26	47.52, 49.01	51.20	50.64, 51.75	< 0.001
C	47.62	46.90, 48.33	51.64	51.08, 52.21	< 0.001
Imp	52.53	51.82, 53.24	48.25	47.69, 48.82	< 0.001
SS	52.98	52.30, 53.65	47.95	47.37, 48.52	< 0.001
Cannabis					
#	1265		620		
N	51.02	50.46, 51.58	47.92	47.19, 48.65	< 0.001
E	49.69	49.12, 50.27	50.63	49.91, 51.34	0.045
O	52.47	51.96, 52.99	44.95	44.20, 45.70	< 0.001
A	48.81	48.25, 49.37	52.42	51.69, 53.15	< 0.001
C	48.08	47.53, 48.63	53.92	53.22, 54.61	< 0.001
Imp	52.05	51.50, 52.61	45.81	45.13, 46.49	< 0.001
SS	52.95	52.44, 53.45	43.99	43.28, 44.70	< 0.001
Chocolate					
#	1850		35		
N	50.00	49.54, 50.45	50.21	46.31, 54.11	0.912

Continued on the next page

Table C.1. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
E	49.99	49.53, 50.44	50.75	47.67, 53.84	0.621
O	50.05	49.59, 50.51	47.41	44.43, 50.38	0.083
A	50.03	49.57, 50.48	48.61	44.55, 52.67	0.486
C	49.97	49.52, 50.43	51.55	47.83, 55.27	0.399
Imp	49.98	49.53, 50.44	50.81	47.30, 54.31	0.640
SS	50.01	49.55, 50.46	49.51	45.65, 53.37	0.796
Cocaine					
#	687		1198		
N	51.78	51.02, 52.54	48.98	48.42, 49.53	< 0.001
E	50.27	49.49, 51.05	49.84	49.29, 50.40	0.381
O	52.58	51.88, 53.29	48.52	47.95, 49.09	< 0.001
A	47.69	46.91, 48.47	51.33	50.79, 51.86	< 0.001
C	47.41	46.68, 48.14	51.48	50.93, 52.04	< 0.001
Imp	53.29	52.54, 54.03	48.12	47.57, 48.66	< 0.001
SS	54.40	53.75, 55.06	47.47	46.92, 48.03	< 0.001
Caffeine					
#	1848		37		
N	50.00	49.54, 50.45	50.06	46.57, 53.54	0.974
E	50.07	49.61, 50.52	46.67	43.77, 49.58	0.025
O	50.11	49.65, 50.56	44.65	41.70, 47.61	0.001
A	49.98	49.52, 50.44	51.05	47.67, 54.42	0.529
C	49.93	49.47, 50.39	53.56	50.75, 56.37	0.014
Imp	50.08	49.62, 50.53	46.10	43.15, 49.05	0.010
SS	50.12	49.66, 50.57	44.21	41.16, 47.27	< 0.001
Crack					

Continued on the next page

Table C.1. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
#	191		1694		
N	53.00	51.56, 54.43	49.66	49.19, 50.14	< 0.001
E	48.74	47.28, 50.20	50.14	49.67, 50.62	0.073
O	52.92	51.61, 54.22	49.67	49.19, 50.15	< 0.001
A	47.00	45.44, 48.55	50.34	49.87, 50.81	< 0.001
C	46.14	44.65, 47.63	50.44	49.97, 50.90	< 0.001
Imp	55.48	54.12, 56.85	49.38	48.91, 49.85	< 0.001
SS	55.43	54.26, 56.61	49.39	48.91, 49.86	< 0.001
Ecstasy					
#	751		1134		
N	51.24	50.52, 51.96	49.18	48.60, 49.75	< 0.001
E	50.49	49.74, 51.24	49.68	49.11, 50.24	0.089
O	53.61	52.95, 54.26	47.61	47.04, 48.18	< 0.001
A	48.47	47.73, 49.20	51.01	50.45, 51.58	< 0.001
C	47.22	46.52, 47.93	51.84	51.28, 52.40	< 0.001
Imp	53.04	52.35, 53.74	47.98	47.42, 48.55	< 0.001
SS	54.97	54.37, 55.56	46.71	46.15, 47.27	< 0.001
Heroin					
#	212		1673		
N	54.53	53.21, 55.85	49.43	48.95, 49.90	< 0.001
E	48.38	46.90, 49.85	50.21	49.73, 50.68	0.021
O	54.24	53.03, 55.45	49.46	48.98, 49.94	< 0.001
A	45.51	43.98, 47.04	50.57	50.10, 51.03	< 0.001
C	45.81	44.45, 47.17	50.53	50.06, 51.00	< 0.001
Imp	55.74	54.43, 57.06	49.27	48.80, 49.74	< 0.001

Continued on the next page

Table C.1. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
SS	56.02	54.84, 57.20	49.24	48.76, 49.71	< 0.001
Ketamine					
#	350		1535		
N	51.35	50.34, 52.36	49.69	49.19, 50.20	0.004
E	50.31	49.19, 51.43	49.93	49.44, 50.42	0.542
O	53.85	52.88, 54.82	49.12	48.62, 49.62	< 0.001
A	47.78	46.64, 48.91	50.51	50.02, 51.00	< 0.001
C	46.79	45.74, 47.85	50.73	50.24, 51.22	< 0.001
Imp	53.81	52.77, 54.85	49.13	48.64, 49.62	< 0.001
SS	55.25	54.40, 56.11	48.80	48.30, 49.30	< 0.001
Legal highs					
#	762		1123		
N	51.43	50.71, 52.16	49.03	48.45, 49.60	< 0.001
E	49.66	48.89, 50.42	50.23	49.68, 50.79	0.228
O	54.29	53.67, 54.91	47.09	46.52, 47.66	< 0.001
A	48.59	47.84, 49.34	50.96	50.40, 51.51	< 0.001
C	46.92	46.20, 47.65	52.09	51.54, 52.63	< 0.001
Imp	53.19	52.49, 53.89	47.83	47.28, 48.39	< 0.001
SS	55.14	54.55, 55.72	46.51	45.95, 47.08	< 0.001
LSD					
#	557		1328		
N	50.80	49.98, 51.62	49.66	49.12, 50.20	0.023
E	50.01	49.11, 50.92	49.99	49.48, 50.51	0.971
O	55.23	54.52, 55.93	47.81	47.28, 48.33	< 0.001
A	48.40	47.53, 49.28	50.67	50.15, 51.19	< 0.001

Continued on the next page

Table C.1. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
C	47.46	46.64, 48.29	51.06	50.53, 51.59	< 0.001
Imp	53.47	52.69, 54.25	48.54	48.01, 49.08	< 0.001
SS	55.62	54.95, 56.30	47.64	47.11, 48.17	< 0.001
Methadone					
#	417		1468		
N	53.33	52.38, 54.28	49.05	48.55, 49.56	< 0.001
E	47.93	46.85, 49.01	50.59	50.10, 51.08	< 0.001
O	53.78	52.87, 54.68	48.93	48.42, 49.43	< 0.001
A	47.06	46.03, 48.09	50.84	50.34, 51.33	< 0.001
C	46.16	45.17, 47.15	51.09	50.60, 51.58	< 0.001
Imp	53.64	52.70, 54.59	48.97	48.46, 49.47	< 0.001
SS	54.45	53.56, 55.33	48.74	48.23, 49.24	< 0.001
Magic mushrooms					
#	694		1191		
N	50.66	49.91, 51.40	49.62	49.05, 50.19	0.029
E	50.10	49.29, 50.90	49.94	49.40, 50.49	0.758
O	54.34	53.70, 54.99	47.47	46.91, 48.03	< 0.001
A	48.52	47.75, 49.30	50.86	50.31, 51.41	< 0.001
C	47.46	46.73, 48.20	51.48	50.92, 52.04	< 0.001
Imp	53.24	52.53, 53.95	48.11	47.55, 48.67	< 0.001
SS	54.78	54.18, 55.39	47.21	46.65, 47.78	< 0.001
Nicotine					
#	1264		621		
N	50.90	50.35, 51.45	48.16	47.40, 48.93	< 0.001
E	49.91	49.35, 50.48	50.18	49.43, 50.93	0.582

Continued on the next page

Table C.1. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
O	51.46	50.91, 52.00	47.04	46.28, 47.79	< 0.001
A	49.18	48.63, 49.73	51.67	50.90, 52.44	< 0.001
C	48.58	48.02, 49.14	52.89	52.17, 53.61	< 0.001
Imp	51.53	50.97, 52.09	46.89	46.17, 47.60	< 0.001
SS	52.13	51.60, 52.67	45.66	44.93, 46.39	< 0.001
VSA					
#	230		1655		
N	52.80	51.49, 54.12	49.61	49.13, 50.09	< 0.001
E	48.91	47.41, 50.41	50.15	49.68, 50.62	0.122
O	54.21	53.01, 55.40	49.42	48.93, 49.90	< 0.001
A	47.28	45.90, 48.67	50.38	49.90, 50.85	< 0.001
C	45.13	43.79, 46.47	50.68	50.21, 51.15	< 0.001
Imp	55.10	53.90, 56.30	49.29	48.81, 49.77	< 0.001
SS	56.76	55.73, 57.78	49.06	48.58, 49.54	< 0.001
Heroin pleiad					
#	832		1053		
N	51.65	50.96, 52.35	48.70	48.11, 49.28	< 0.001
E	49.70	48.98, 50.41	50.24	49.66, 50.82	0.246
O	52.75	52.11, 53.39	47.83	47.23, 48.43	< 0.001
A	47.91	47.20, 48.62	51.65	51.08, 52.21	< 0.001
C	47.37	46.69, 48.05	52.08	51.50, 52.65	< 0.001
Imp	52.93	52.26, 53.60	47.68	47.11, 48.26	< 0.001
SS	54.00	53.39, 54.60	46.84	46.26, 47.43	< 0.001
Ecstasy pleiad					
#	1317		568		

Continued on the next page

Table C.1. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
N	51.00	50.44, 51.55	47.68	46.94, 48.42	< 0.001
E	49.69	49.13, 50.26	50.71	49.98, 51.44	0.031
O	52.31	51.81, 52.82	44.64	43.86, 45.42	< 0.001
A	48.75	48.20, 49.30	52.89	52.15, 53.64	< 0.001
C	48.10	47.56, 48.64	54.40	53.70, 55.10	< 0.001
Imp	51.98	51.44, 52.52	45.41	44.72, 46.11	< 0.001
SS	52.83	52.34, 53.33	43.43	42.71, 44.14	< 0.001
Benzodiazepines pleiad					
#	1089		796		
N	51.55	50.94, 52.15	47.88	47.23, 48.54	< 0.001
E	49.63	49.01, 50.25	50.51	49.85, 51.16	0.057
O	52.28	51.70, 52.85	46.89	46.22, 47.55	< 0.001
A	48.57	47.96, 49.18	51.96	51.31, 52.60	< 0.001
C	47.90	47.30, 48.50	52.87	52.23, 53.51	< 0.001
Imp	52.16	51.56, 52.75	47.05	46.40, 47.69	< 0.001
SS	53.00	52.44, 53.55	45.90	45.24, 46.55	< 0.001
Illicit drugs					
#	1418		467		
N	51.05	50.52, 51.58	46.80	46.02, 47.58	< 0.001
E	49.54	49.00, 50.09	51.39	50.62, 52.16	< 0.001
O	51.77	51.27, 52.28	44.61	43.77, 45.45	< 0.001
A	48.98	48.45, 49.51	53.11	52.31, 53.91	< 0.001
C	48.37	47.84, 48.89	54.96	54.24, 55.69	< 0.001
Imp	51.64	51.12, 52.16	45.02	44.27, 45.76	< 0.001
SS	52.19	51.70, 52.68	43.35	42.56, 44.13	< 0.001

Table C.2. Mean T-$score_{sample}$ (MT) and 95% CI for it for groups of users and non-users with year-based definition of users

Factor	Users		Non-users		p-value
	MT	95% CI	MT	95% CI	
Alcohol					
#	1749		136		
N	49.96	49.49, 50.43	50.55	48.87, 52.23	0.501
E	50.12	49.64, 50.59	48.52	46.89, 50.14	0.064
O	50.07	49.60, 50.54	49.11	47.46, 50.76	0.271
A	49.95	49.48, 50.42	50.70	49.03, 52.37	0.392
C	49.94	49.47, 50.41	50.73	49.03, 52.44	0.378
Imp	50.08	49.61, 50.55	48.96	47.31, 50.62	0.203
SS	50.24	49.77, 50.71	46.91	45.24, 48.58	< 0.001
Amphetamine					
#	436		1449		
N	52.40	51.43, 53.36	49.28	48.77, 49.78	< 0.001
E	49.48	48.45, 50.50	50.16	49.66, 50.66	0.243
O	53.83	52.92, 54.74	48.85	48.34, 49.35	< 0.001
A	47.47	46.45, 48.48	50.76	50.27, 51.26	< 0.001
C	45.91	44.94, 46.88	51.23	50.74, 51.72	< 0.001
Imp	54.93	54.02, 55.84	48.52	48.02, 49.01	< 0.001
SS	55.79	55.01, 56.56	48.26	47.75, 48.76	< 0.001
Amyl nitrite					
#	133		1752		
N	51.57	49.85, 53.29	49.88	49.41, 50.35	0.063
E	50.25	48.44, 52.06	49.98	49.51, 50.45	0.778
O	51.97	50.36, 53.59	49.85	49.38, 50.32	0.014
A	46.35	44.64, 48.07	50.28	49.81, 50.74	< 0.001

Continued on the next page

Table C.2. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
C	46.92	45.25, 48.60	50.23	49.77, 50.70	< 0.001
Imp	52.39	50.71, 54.08	49.82	49.35, 50.29	0.004
SS	55.06	53.67, 56.44	49.62	49.15, 50.09	< 0.001
Benzodiazepines					
#	535		1350		
N	53.69	52.84, 54.54	48.54	48.03, 49.05	< 0.001
E	48.91	47.97, 49.85	50.43	49.92, 50.94	0.005
O	53.00	52.17, 53.83	48.81	48.28, 49.34	< 0.001
A	47.55	46.64, 48.46	50.97	50.46, 51.48	< 0.001
C	47.07	46.19, 47.94	51.16	50.65, 51.68	< 0.001
Imp	53.35	52.52, 54.19	48.67	48.15, 49.19	< 0.001
SS	54.10	53.32, 54.87	48.38	47.85, 48.90	< 0.001
Cannabis					
#	999		886		
N	51.10	50.46, 51.74	48.76	48.14, 49.38	< 0.001
E	49.73	49.08, 50.39	50.30	49.69, 50.91	0.217
O	53.68	53.13, 54.23	45.85	45.22, 46.48	< 0.001
A	48.77	48.13, 49.40	51.39	50.76, 52.02	< 0.001
C	47.33	46.70, 47.95	53.02	52.42, 53.62	< 0.001
Imp	52.74	52.13, 53.36	46.91	46.30, 47.51	< 0.001
SS	54.29	53.76, 54.81	45.17	44.55, 45.79	< 0.001
Chocolate					
#	1840		45		
N	50.01	49.56, 50.47	49.45	46.05, 52.86	0.744
E	49.99	49.53, 50.45	50.49	47.75, 53.24	0.716

Continued on the next page

Table C.2. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
O	50.04	49.58, 50.50	48.27	45.59, 50.94	0.194
A	50.02	49.56, 50.48	49.17	45.72, 52.62	0.626
C	49.97	49.51, 50.42	51.41	47.81, 55.01	0.425
Imp	49.98	49.52, 50.43	50.94	47.98, 53.90	0.520
SS	50.00	49.54, 50.45	50.06	46.84, 53.29	0.968
Cocaine					
#	417		1468		
N	52.16	51.19, 53.13	49.39	48.88, 49.89	< 0.001
E	50.93	49.88, 51.98	49.74	49.24, 50.23	0.044
O	52.79	51.87, 53.71	49.21	48.70, 49.72	< 0.001
A	46.84	45.81, 47.86	50.90	50.41, 51.39	< 0.001
C	46.81	45.86, 47.76	50.91	50.40, 51.41	< 0.001
Imp	54.29	53.33, 55.25	48.78	48.29, 49.28	< 0.001
SS	55.63	54.81, 56.45	48.40	47.90, 48.90	< 0.001
Caffeine					
#	1824		61		
N	49.99	49.53, 50.45	50.36	47.75, 52.96	0.783
E	50.10	49.64, 50.56	46.92	44.34, 49.50	0.018
O	50.10	49.64, 50.56	47.07	44.45, 49.69	0.026
A	49.96	49.50, 50.42	51.15	48.61, 53.69	0.360
C	49.91	49.45, 50.37	52.78	50.14, 55.43	0.036
Imp	50.13	49.67, 50.59	46.23	43.95, 48.51	0.001
SS	50.18	49.72, 50.63	44.70	42.22, 47.18	< 0.001
Crack					
#	79		1806		

Continued on the next page

Table C.2. *Continued*

Factor	Users		Non-users		p-value
	MT	95% CI	MT	95% CI	
N	54.06	51.77, 56.35	49.82	49.36, 50.28	0.001
E	49.24	47.20, 51.28	50.03	49.57, 50.50	0.455
O	52.90	50.72, 55.08	49.87	49.41, 50.33	0.008
A	46.65	44.18, 49.12	50.15	49.69, 50.60	0.007
C	45.08	42.92, 47.25	50.22	49.76, 50.67	< 0.001
Imp	56.55	54.62, 58.48	49.71	49.25, 50.17	< 0.001
SS	56.56	54.89, 58.22	49.71	49.25, 50.17	< 0.001
Ecstasy					
#	517		1368		
N	50.71	49.82, 51.61	49.73	49.21, 50.25	0.062
E	51.51	50.60, 52.42	49.43	48.91, 49.95	< 0.001
O	54.10	53.32, 54.89	48.45	47.93, 48.97	< 0.001
A	48.55	47.64, 49.45	50.55	50.03, 51.07	< 0.001
C	47.02	46.17, 47.87	51.13	50.61, 51.65	< 0.001
Imp	53.74	52.91, 54.57	48.59	48.07, 49.11	< 0.001
SS	55.85	55.16, 56.53	47.79	47.27, 48.31	< 0.001
Heroin					
#	118		1767		
N	55.37	53.62, 57.12	49.64	49.18, 50.10	< 0.001
E	47.69	45.68, 49.69	50.15	49.69, 50.62	0.019
O	53.45	51.80, 55.11	49.77	49.30, 50.24	< 0.001
A	44.38	42.26, 46.50	50.38	49.92, 50.83	< 0.001
C	45.46	43.52, 47.39	50.30	49.84, 50.76	< 0.001
Imp	56.01	54.29, 57.74	49.60	49.14, 50.06	< 0.001
SS	56.41	54.81, 58.00	49.57	49.11, 50.04	< 0.001

Continued on the next page

Table C.2. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
Ketamine					
#	208		1677		
N	51.52	50.20, 52.84	49.81	49.33, 50.29	0.017
E	51.02	49.48, 52.57	49.87	49.40, 50.34	0.162
O	54.18	52.88, 55.48	49.48	49.01, 49.96	< 0.001
A	47.90	46.42, 49.38	50.26	49.79, 50.73	0.003
C	46.34	44.91, 47.76	50.45	49.98, 50.93	< 0.001
Imp	54.18	52.83, 55.53	49.48	49.01, 49.96	< 0.001
SS	55.93	54.83, 57.02	49.26	48.79, 49.74	< 0.001
Legal highs					
#	564		1321		
N	51.27	50.41, 52.12	49.46	48.93, 49.99	< 0.001
E	50.02	49.11, 50.92	49.99	49.48, 50.51	0.966
O	54.49	53.76, 55.22	48.08	47.55, 48.62	< 0.001
A	48.10	47.23, 48.96	50.81	50.29, 51.34	< 0.001
C	46.56	45.71, 47.41	51.47	50.95, 51.98	< 0.001
Imp	53.82	53.00, 54.63	48.37	47.85, 48.89	< 0.001
SS	56.00	55.35, 56.64	47.44	46.92, 47.97	< 0.001
LSD					
#	380		1505		
N	49.98	48.97, 50.99	50.00	49.50, 50.51	0.969
E	50.72	49.63, 51.81	49.82	49.32, 50.31	0.140
O	56.29	55.48, 57.10	48.41	47.91, 48.91	< 0.001
A	49.05	47.98, 50.12	50.24	49.74, 50.74	0.048
C	47.74	46.71, 48.77	50.57	50.07, 51.07	< 0.001

Continued on the next page

Table C.2. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
Imp	53.50	52.56, 54.43	49.12	48.61, 49.62	< 0.001
SS	56.34	55.56, 57.12	48.40	47.90, 48.90	< 0.001
Methadone					
#	320		1565		
N	53.74	52.65, 54.84	49.23	48.75, 49.72	< 0.001
E	47.75	46.47, 49.03	50.46	49.99, 50.94	< 0.001
O	53.81	52.76, 54.86	49.22	48.73, 49.71	< 0.001
A	46.53	45.33, 47.73	50.71	50.23, 51.19	< 0.001
C	46.01	44.87, 47.15	50.82	50.33, 51.30	< 0.001
Imp	53.77	52.67, 54.87	49.23	48.74, 49.72	< 0.001
SS	54.73	53.73, 55.74	49.03	48.54, 49.52	< 0.001
Magic mushrooms					
#	434		1451		
N	50.33	49.40, 51.26	49.90	49.38, 50.42	0.431
E	50.71	49.66, 51.76	49.79	49.29, 50.28	0.118
O	55.72	54.95, 56.48	48.29	47.78, 48.80	< 0.001
A	48.51	47.48, 49.53	50.45	49.95, 50.95	0.001
C	47.30	46.36, 48.24	50.81	50.30, 51.32	< 0.001
Imp	53.74	52.86, 54.63	48.88	48.37, 49.39	< 0.001
SS	55.92	55.19, 56.64	48.23	47.72, 48.74	< 0.001
Nicotine					
#	1060		825		
N	51.16	50.55, 51.76	48.52	47.85, 49.18	< 0.001
E	49.80	49.17, 50.42	50.26	49.61, 50.91	0.309
O	51.91	51.31, 52.51	47.55	46.90, 48.20	< 0.001

Continued on the next page

Table C.2. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
A	49.04	48.42, 49.65	51.24	50.58, 51.90	< 0.001
C	47.98	47.38, 48.57	52.60	51.95, 53.25	< 0.001
Imp	52.08	51.46, 52.69	47.33	46.71, 47.95	< 0.001
SS	52.83	52.26, 53.41	46.36	45.72, 47.00	< 0.001
VSA					
#	95		1790		
N	53.81	51.67, 55.95	49.80	49.34, 50.26	< 0.001
E	49.55	47.32, 51.79	50.02	49.56, 50.48	0.683
O	53.59	51.58, 55.60	49.81	49.35, 50.27	< 0.001
A	47.48	45.35, 49.61	50.13	49.67, 50.60	0.017
C	45.31	43.18, 47.44	50.25	49.79, 50.71	< 0.001
Imp	54.61	52.73, 56.49	49.76	49.29, 50.22	< 0.001
SS	56.24	54.48, 57.99	49.67	49.21, 50.13	< 0.001
Heroin pleiad					
#	585		1300		
N	52.24	51.41, 53.07	48.99	48.46, 49.52	< 0.001
E	49.59	48.68, 50.49	50.19	49.67, 50.70	0.260
O	53.02	52.25, 53.79	48.64	48.10, 49.18	< 0.001
A	47.10	46.24, 47.97	51.30	50.79, 51.81	< 0.001
C	46.72	45.91, 47.53	51.47	50.95, 52.00	< 0.001
Imp	53.75	52.94, 54.56	48.31	47.79, 48.83	< 0.001
SS	54.87	54.16, 55.59	47.81	47.28, 48.34	< 0.001
Ecstasy pleiad					
#	1089		796		
N	51.16	50.54, 51.78	48.41	47.77, 49.05	< 0.001

Continued on the next page

Table C.2. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
E	49.59	48.96, 50.22	50.56	49.93, 51.20	0.033
O	53.32	52.78, 53.85	45.46	44.81, 46.12	< 0.001
A	48.55	47.95, 49.16	51.98	51.33, 52.63	< 0.001
C	47.43	46.83, 48.02	53.52	52.91, 54.13	< 0.001
Imp	52.70	52.11, 53.29	46.31	45.69, 46.92	< 0.001
SS	54.00	53.49, 54.51	44.53	43.89, 45.17	< 0.001
Benzodiazepines pleiad					
#	830		1055		
N	52.23	51.53, 52.93	48.25	47.68, 48.82	< 0.001
E	49.45	48.71, 50.18	50.43	49.87, 51.00	0.037
O	52.86	52.20, 53.51	47.75	47.16, 48.34	< 0.001
A	47.88	47.16, 48.59	51.67	51.11, 52.23	< 0.001
C	47.21	46.51, 47.90	52.20	51.64, 52.76	< 0.001
Imp	53.09	52.41, 53.76	47.57	47.00, 48.14	< 0.001
SS	54.13	53.52, 54.74	46.75	46.17, 47.33	< 0.001
Illicit drugs					
#	1179		706		
N	51.38	50.79, 51.97	48.19	47.53, 48.84	< 0.001
E	49.40	48.80, 50.01	50.93	50.27, 51.59	0.003
O	52.84	52.31, 53.37	45.64	44.95, 46.33	< 0.001
A	48.59	48.00, 49.17	51.99	51.34, 52.65	< 0.001
C	47.61	47.04, 48.19	53.63	52.97, 54.28	< 0.001
Imp	52.46	51.89, 53.03	46.10	45.46, 46.74	< 0.001
SS	53.48	52.97, 53.98	44.23	43.55, 44.90	< 0.001

Table C.3. Mean T-$score_{sample}$ (MT) and 95% CI for it for groups of users and non-users with month-based definition of users

Factor	Users		Non-users		p-value
	MT	95% CI	MT	95% CI	
Alcohol					
#	1551		334		
N	49.85	49.35, 50.35	50.69	49.61, 51.78	0.167
E	50.62	50.12, 51.11	47.14	46.11, 48.17	< 0.001
O	50.22	49.73, 50.72	48.96	47.87, 50.06	0.040
A	50.10	49.60, 50.60	49.55	48.50, 50.60	0.357
C	50.15	49.66, 50.65	49.28	48.22, 50.35	0.146
Imp	50.06	49.57, 50.56	49.71	48.61, 50.82	0.571
SS	50.34	49.85, 50.84	48.41	47.32, 49.50	0.002
Amphetamine					
#	238		1647		
N	52.78	51.49, 54.07	49.60	49.12, 50.08	< 0.001
E	49.07	47.63, 50.51	50.13	49.66, 50.61	0.166
O	52.94	51.69, 54.19	49.57	49.09, 50.06	< 0.001
A	46.57	45.21, 47.92	50.50	50.02, 50.97	< 0.001
C	45.06	43.75, 46.37	50.71	50.24, 51.19	< 0.001
Imp	55.57	54.32, 56.83	49.19	48.72, 49.67	< 0.001
SS	55.37	54.23, 56.51	49.22	48.75, 49.70	< 0.001
Amyl nitrite					
#	41		1844		
N	49.37	46.12, 52.61	50.01	49.56, 50.47	0.691
E	49.83	46.38, 53.28	50.00	49.55, 50.46	0.922
O	50.40	47.06, 53.73	49.99	49.53, 50.45	0.808
A	45.43	41.87, 49.00	50.10	49.65, 50.56	0.012
C	47.31	44.61, 50.01	50.06	49.60, 50.52	0.049

Continued on the next page

Table C.3. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
Imp	54.16	50.89, 57.43	49.91	49.45, 50.36	0.013
SS	57.33	54.76, 59.89	49.84	49.38, 50.29	< 0.001
Benzodiazepines					
#	299		1586		
N	55.26	54.11, 56.41	49.01	48.53, 49.48	< 0.001
E	48.06	46.79, 49.33	50.37	49.89, 50.84	0.001
O	52.68	51.58, 53.77	49.49	49.00, 49.99	< 0.001
A	46.72	45.47, 47.97	50.62	50.14, 51.09	< 0.001
C	46.54	45.40, 47.69	50.65	50.17, 51.14	< 0.001
Imp	54.06	52.96, 55.16	49.23	48.75, 49.72	< 0.001
SS	54.36	53.34, 55.38	49.18	48.69, 49.67	< 0.001
Cannabis					
#	788		1097		
N	50.72	49.99, 51.45	49.48	48.91, 50.06	0.009
E	50.06	49.31, 50.80	49.96	49.40, 50.52	0.839
O	54.34	53.74, 54.94	46.88	46.30, 47.46	< 0.001
A	48.66	47.93, 49.38	50.97	50.39, 51.54	< 0.001
C	47.26	46.55, 47.97	51.97	51.41, 52.52	< 0.001
Imp	53.14	52.46, 53.82	47.74	47.17, 48.31	< 0.001
SS	54.82	54.25, 55.40	46.53	45.96, 47.11	< 0.001
Chocolate					
#	1786		99		
N	49.98	49.52, 50.44	50.40	48.19, 52.60	0.714
E	50.03	49.57, 50.49	49.43	47.38, 51.49	0.574
O	50.05	49.58, 50.51	49.13	47.37, 50.90	0.323

Continued on the next page

Table C.3. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
A	50.04	49.57, 50.50	49.36	47.19, 51.53	0.548
C	49.95	49.49, 50.41	50.89	48.82, 52.97	0.380
Imp	49.97	49.50, 50.43	50.61	48.75, 52.46	0.508
SS	50.02	49.55, 50.48	49.72	47.63, 51.81	0.783
Cocaine					
#	159		1726		
N	52.87	51.29, 54.46	49.74	49.27, 50.21	< 0.001
E	51.34	49.50, 53.18	49.88	49.41, 50.34	0.129
O	52.53	51.00, 54.06	49.77	49.30, 50.24	0.001
A	45.75	44.05, 47.46	50.39	49.93, 50.86	< 0.001
C	47.23	45.72, 48.75	50.25	49.78, 50.73	< 0.001
Imp	54.59	53.06, 56.11	49.58	49.11, 50.05	< 0.001
SS	56.48	55.10, 57.86	49.40	48.94, 49.87	< 0.001
Caffeine					
#	1764		121		
N	50.05	49.59, 50.52	49.21	47.37, 51.05	0.379
E	50.16	49.69, 50.63	47.68	45.94, 49.43	0.008
O	50.12	49.65, 50.58	48.29	46.47, 50.11	0.056
A	49.98	49.51, 50.45	50.27	48.50, 52.04	0.754
C	49.95	49.49, 50.42	50.67	48.77, 52.56	0.473
Imp	50.22	49.75, 50.68	46.85	45.15, 48.55	< 0.001
SS	50.20	49.73, 50.66	47.09	45.29, 48.90	0.001
Crack					
#	20		1865		
N	57.86	52.26, 63.45	49.92	49.46, 50.37	0.008

Continued on the next page

Table C.3. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
E	45.97	41.31, 50.64	50.04	49.59, 50.50	0.085
O	50.89	47.21, 54.57	49.99	49.54, 50.45	0.616
A	42.99	36.35, 49.62	50.08	49.62, 50.53	0.038
C	45.14	40.37, 49.91	50.05	49.60, 50.51	0.045
Imp	55.89	51.51, 60.27	49.94	49.48, 50.39	0.011
SS	57.01	54.05, 59.96	49.92	49.47, 50.38	< 0.001
Ecstasy					
#	240		1645		
N	49.53	48.18, 50.89	50.07	49.59, 50.55	0.465
E	52.24	50.82, 53.65	49.67	49.20, 50.15	0.001
O	54.41	53.23, 55.59	49.36	48.88, 49.84	< 0.001
A	48.10	46.75, 49.45	50.28	49.80, 50.76	0.003
C	47.27	45.96, 48.59	50.40	49.92, 50.88	< 0.001
Imp	53.43	52.19, 54.68	49.50	49.02, 49.98	< 0.001
SS	55.63	54.62, 56.64	49.18	48.69, 49.66	< 0.001
Heroin					
#	53		1832		
N	56.69	54.05, 59.34	49.81	49.35, 50.26	< 0.001
E	45.58	42.06, 49.10	50.13	49.67, 50.58	0.013
O	52.48	49.63, 55.34	49.93	49.47, 50.39	0.082
A	42.18	39.00, 45.35	50.23	49.77, 50.68	< 0.001
C	43.36	40.35, 46.37	50.19	49.74, 50.65	< 0.001
Imp	57.08	54.63, 59.53	49.80	49.34, 50.25	< 0.001
SS	57.30	54.94, 59.66	49.79	49.33, 50.25	< 0.001
Ketamine					

Continued on the next page

Table C.3. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
#	79		1806		
N	51.29	49.14, 53.45	49.94	49.48, 50.41	0.227
E	49.62	46.74, 52.49	50.02	49.56, 50.47	0.785
O	54.79	52.59, 56.98	49.79	49.33, 50.25	< 0.001
A	46.90	44.15, 49.66	50.14	49.68, 50.59	0.024
C	45.03	42.50, 47.56	50.22	49.76, 50.67	< 0.001
Imp	53.39	51.39, 55.38	49.85	49.39, 50.31	0.001
SS	54.96	53.13, 56.80	49.78	49.32, 50.25	< 0.001
Legal highs					
#	241		1644		
N	52.02	50.68, 53.36	49.70	49.23, 50.18	0.002
E	49.10	47.59, 50.61	50.13	49.66, 50.60	0.200
O	54.37	53.22, 55.53	49.36	48.88, 49.84	< 0.001
A	46.83	45.50, 48.16	50.46	49.99, 50.94	< 0.001
C	45.30	44.01, 46.60	50.69	50.21, 51.16	< 0.001
Imp	54.21	53.00, 55.41	49.38	48.90, 49.86	< 0.001
SS	56.05	55.05, 57.04	49.11	48.63, 49.60	< 0.001
LSD					
#	166		1719		
N	50.55	48.97, 52.12	49.95	49.48, 50.42	0.472
E	51.28	49.53, 53.04	49.88	49.41, 50.34	0.128
O	57.28	56.15, 58.41	49.30	48.83, 49.77	< 0.001
A	48.92	47.35, 50.48	50.10	49.63, 50.58	0.153
C	47.10	45.59, 48.60	50.28	49.81, 50.75	< 0.001
Imp	53.35	51.90, 54.81	49.68	49.20, 50.15	< 0.001

Continued on the next page

Table C.3. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
SS	56.50	55.32, 57.68	49.37	48.90, 49.84	< 0.001
Methadone					
#	171		1714		
N	54.53	53.00, 56.06	49.55	49.08, 50.02	< 0.001
E	46.86	45.08, 48.65	50.31	49.85, 50.78	< 0.001
O	52.89	51.37, 54.40	49.71	49.24, 50.18	< 0.001
A	46.18	44.50, 47.87	50.38	49.92, 50.85	< 0.001
C	45.44	43.83, 47.06	50.45	49.99, 50.92	< 0.001
Imp	53.89	52.36, 55.42	49.61	49.14, 50.08	< 0.001
SS	54.71	53.25, 56.17	49.53	49.06, 50.00	< 0.001
Magic mushrooms					
#	159		1726		
N	49.91	48.41, 51.42	50.01	49.53, 50.48	0.906
E	50.31	48.50, 52.12	49.97	49.51, 50.44	0.720
O	56.92	55.77, 58.07	49.36	48.89, 49.83	< 0.001
A	48.57	46.94, 50.19	50.13	49.66, 50.60	0.070
C	46.85	45.25, 48.45	50.29	49.82, 50.76	< 0.001
Imp	53.61	52.13, 55.09	49.67	49.20, 50.14	< 0.001
SS	56.18	54.94, 57.41	49.43	48.96, 49.90	< 0.001
Nicotine					
#	875		1010		
N	51.11	50.43, 51.79	49.04	48.44, 49.64	< 0.001
E	49.98	49.29, 50.66	50.02	49.42, 50.62	0.924
O	51.86	51.19, 52.52	48.39	47.79, 48.99	< 0.001
A	49.02	48.34, 49.71	50.85	50.25, 51.44	< 0.001

Continued on the next page

Table C.3. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
C	47.69	47.03, 48.34	52.00	51.41, 52.60	< 0.001
Imp	52.48	51.80, 53.15	47.86	47.28, 48.43	< 0.001
SS	52.88	52.24, 53.51	47.51	46.91, 48.10	< 0.001
VSA					
#	34		1851		
N	51.34	47.49, 55.20	49.98	49.52, 50.43	0.479
E	51.80	48.00, 55.60	49.97	49.51, 50.42	0.338
O	54.65	51.65, 57.65	49.91	49.46, 50.37	0.003
A	45.91	42.12, 49.71	50.08	49.62, 50.53	0.034
C	47.22	43.63, 50.81	50.05	49.60, 50.51	0.121
Imp	55.93	53.25, 58.62	49.89	49.43, 50.35	< 0.001
SS	58.61	56.27, 60.95	49.84	49.39, 50.30	< 0.001
Heroin pleiad					
#	309		1576		
N	53.20	52.08, 54.32	49.37	48.88, 49.86	< 0.001
E	48.83	47.49, 50.18	50.23	49.76, 50.70	0.054
O	52.60	51.51, 53.69	49.49	49.00, 49.98	< 0.001
A	46.40	45.22, 47.59	50.71	50.22, 51.19	< 0.001
C	46.69	45.54, 47.84	50.65	50.16, 51.13	< 0.001
Imp	54.24	53.12, 55.35	49.17	48.69, 49.65	< 0.001
SS	55.31	54.27, 56.36	48.96	48.47, 49.44	< 0.001
Ecstasy pleiad					
#	921		964		
N	50.98	50.30, 51.66	49.06	48.47, 49.66	< 0.001
E	49.72	49.02, 50.41	50.27	49.69, 50.85	0.234

Continued on the next page

Table C.3. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
O	53.86	53.29, 54.43	46.31	45.70, 46.92	< 0.001
A	48.32	47.65, 48.99	51.61	51.01, 52.20	< 0.001
C	47.08	46.42, 47.74	52.79	52.23, 53.35	< 0.001
Imp	53.25	52.61, 53.88	46.90	46.32, 47.48	< 0.001
SS	54.58	54.03, 55.14	45.62	45.03, 46.21	< 0.001
Benzodiazepines pleiad					
#	528		1357		
N	53.44	52.58, 54.30	48.66	48.15, 49.18	< 0.001
E	48.88	47.92, 49.84	50.44	49.93, 50.94	0.005
O	52.57	51.74, 53.40	49.00	48.47, 49.53	< 0.001
A	47.15	46.26, 48.04	51.11	50.60, 51.62	< 0.001
C	46.46	45.60, 47.32	51.38	50.87, 51.89	< 0.001
Imp	53.90	53.06, 54.75	48.48	47.97, 48.99	< 0.001
SS	54.61	53.84, 55.39	48.20	47.68, 48.73	< 0.001
Illicit drugs					
#	996		889		
N	51.33	50.67, 51.98	48.56	47.96, 49.15	< 0.001
E	49.43	48.76, 50.10	51.03	50.47, 51.59	0.002
O	53.46	52.90, 54.03	45.82	45.20, 46.45	< 0.001
A	48.36	47.72, 49.00	51.48	50.87, 52.10	< 0.001
C	47.22	46.58, 47.86	53.36	52.79, 53.92	< 0.001
Imp	52.97	52.36, 53.58	46.40	45.80, 46.99	< 0.001
SS	54.17	53.63, 54.71	44.78	44.17, 45.40	< 0.001

Table C.4. Mean T-$score_{sample}$ (MT) and 95% CI for it for groups of users and non-users with week-based definition of users

Factor	Users		Non-users		p-value
	MT	95% CI	MT	95% CI	
Alcohol					
#	1264		621		
N	49.82	49.28, 50.36	50.37	49.56, 51.19	0.267
E	50.90	50.35, 51.44	48.17	47.38, 48.97	< 0.001
O	50.08	49.54, 50.63	49.83	49.03, 50.62	0.602
A	50.05	49.50, 50.60	49.89	49.09, 50.69	0.745
C	50.19	49.64, 50.74	49.61	48.81, 50.42	0.244
Imp	50.19	49.64, 50.73	49.62	48.81, 50.43	0.253
SS	50.52	49.97, 51.06	48.95	48.15, 49.75	0.001
Amphetamine					
#	163		1722		
N	52.86	51.27, 54.45	49.73	49.26, 50.20	< 0.001
E	48.50	46.78, 50.21	50.14	49.68, 50.61	0.069
O	52.87	51.27, 54.47	49.73	49.26, 50.20	< 0.001
A	46.77	45.09, 48.45	50.31	49.84, 50.77	< 0.001
C	45.29	43.72, 46.85	50.45	49.98, 50.91	< 0.001
Imp	55.77	54.25, 57.29	49.45	48.99, 49.92	< 0.001
SS	54.47	53.07, 55.88	49.58	49.10, 50.05	< 0.001
Amyl nitrite					
#	17		1868		
N	49.64	43.99, 55.28	50.00	49.55, 50.46	0.892
E	45.85	39.05, 52.64	50.04	49.59, 50.49	0.211
O	49.02	43.03, 55.02	50.01	49.56, 50.46	0.732
A	44.36	38.95, 49.77	50.05	49.60, 50.50	0.041
C	44.81	39.74, 49.89	50.05	49.59, 50.50	0.045

Continued on the next page

Table C.4. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
Imp	56.76	50.49, 63.03	49.94	49.49, 50.39	0.035
SS	57.30	53.19, 61.41	49.93	49.48, 50.39	0.002
Benzodiazepines					
#	179		1706		
N	56.56	55.08, 58.04	49.31	48.85, 49.77	< 0.001
E	46.15	44.61, 47.70	50.40	49.93, 50.87	< 0.001
O	52.18	50.72, 53.63	49.77	49.30, 50.25	0.002
A	46.57	44.94, 48.21	50.36	49.89, 50.83	< 0.001
C	46.20	44.75, 47.65	50.40	49.93, 50.87	< 0.001
Imp	53.89	52.47, 55.31	49.59	49.12, 50.06	< 0.001
SS	53.62	52.28, 54.95	49.62	49.14, 50.10	< 0.001
Cannabis					
#	648		1237		
N	50.62	49.82, 51.42	49.68	49.13, 50.22	0.055
E	50.17	49.36, 50.97	49.91	49.37, 50.46	0.606
O	54.70	54.05, 55.35	47.54	46.99, 48.09	< 0.001
A	48.78	48.00, 49.56	50.64	50.09, 51.19	< 0.001
C	47.45	46.67, 48.23	51.34	50.80, 51.88	< 0.001
Imp	53.02	52.29, 53.75	48.42	47.87, 48.97	< 0.001
SS	54.87	54.22, 55.51	47.45	46.90, 48.00	< 0.001
Chocolate					
#	1490		395		
N	49.95	49.44, 50.45	50.20	49.18, 51.23	0.660
E	50.22	49.72, 50.73	49.16	48.14, 50.17	0.065
O	49.89	49.37, 50.40	50.42	49.48, 51.36	0.327

Continued on the next page

Table C.4. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
A	50.20	49.70, 50.70	49.23	48.19, 50.27	0.099
C	50.13	49.62, 50.63	49.51	48.50, 50.53	0.286
Imp	49.93	49.43, 50.44	50.25	49.26, 51.24	0.580
SS	49.72	49.21, 50.22	51.07	50.06, 52.08	0.019
Cocaine					
#	60		1825		
N	53.24	50.50, 55.98	49.89	49.44, 50.35	0.019
E	52.03	49.22, 54.84	49.93	49.48, 50.39	0.146
O	51.40	48.83, 53.97	49.95	49.49, 50.41	0.273
A	43.74	40.80, 46.68	50.21	49.75, 50.66	< 0.001
C	46.72	44.32, 49.11	50.11	49.65, 50.57	0.007
Imp	54.87	52.42, 57.32	49.84	49.38, 50.30	< 0.001
SS	57.38	55.14, 59.62	49.76	49.30, 50.21	< 0.001
Caffeine					
#	1658		227		
N	50.07	49.59, 50.55	49.46	48.11, 50.81	0.400
E	50.20	49.72, 50.68	48.54	47.27, 49.81	0.016
O	50.04	49.56, 50.52	49.71	48.40, 51.03	0.645
A	49.93	49.45, 50.41	50.48	49.15, 51.81	0.446
C	49.95	49.47, 50.43	50.37	49.01, 51.72	0.569
Imp	50.17	49.69, 50.65	48.76	47.48, 50.04	0.043
SS	50.11	49.63, 50.60	49.16	47.89, 50.43	0.169
Crack					
#	11		1874		
N	55.26	45.53, 64.99	49.97	49.52, 50.42	0.254

Continued on the next page

Table C.4. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
E	46.46	38.43, 54.50	50.02	49.57, 50.47	0.348
O	48.01	43.74, 52.28	50.01	49.56, 50.47	0.324
A	39.62	30.12, 49.11	50.06	49.61, 50.51	0.034
C	44.28	36.73, 51.83	50.03	49.58, 50.49	0.121
Imp	56.94	50.59, 63.29	49.96	49.51, 50.41	0.034
SS	58.03	54.16, 61.91	49.95	49.50, 50.41	0.001
Ecstasy					
#	84		1801		
N	50.28	47.99, 52.58	49.99	49.53, 50.45	0.803
E	53.37	50.71, 56.02	49.84	49.39, 50.30	0.011
O	56.15	54.15, 58.15	49.71	49.25, 50.17	< 0.001
A	48.56	46.38, 50.75	50.07	49.61, 50.53	0.184
C	46.98	44.78, 49.18	50.14	49.68, 50.60	0.006
Imp	55.71	53.49, 57.93	49.73	49.28, 50.19	< 0.001
SS	56.57	54.81, 58.34	49.69	49.23, 50.15	< 0.001
Heroin					
#	29		1856		
N	58.65	55.84, 61.47	49.86	49.41, 50.32	< 0.001
E	44.77	39.81, 49.73	50.08	49.63, 50.53	0.038
O	52.41	48.00, 56.81	49.96	49.51, 50.42	0.268
A	41.48	38.25, 44.70	50.13	49.68, 50.59	< 0.001
C	43.04	38.95, 47.12	50.11	49.66, 50.56	0.001
Imp	56.79	53.59, 59.99	49.89	49.44, 50.35	< 0.001
SS	56.36	52.91, 59.81	49.90	49.45, 50.36	0.001
Ketamine					

Continued on the next page

Table C.4. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
#	37		1848		
N	50.47	46.78, 54.16	49.99	49.54, 50.45	0.795
E	47.23	42.01, 52.46	50.06	49.61, 50.50	0.282
O	54.39	50.87, 57.90	49.91	49.46, 50.37	0.015
A	44.50	40.15, 48.85	50.11	49.66, 50.56	0.013
C	44.99	41.13, 48.85	50.10	49.65, 50.55	0.011
Imp	53.74	50.75, 56.74	49.93	49.47, 50.38	0.015
SS	55.44	52.84, 58.04	49.89	49.43, 50.35	< 0.001
Legal highs					
#	131		1754		
N	53.13	51.32, 54.94	49.77	49.30, 50.23	0.001
E	47.12	45.06, 49.18	50.22	49.76, 50.67	0.004
O	53.16	51.53, 54.79	49.76	49.30, 50.23	< 0.001
A	46.25	44.62, 47.87	50.28	49.81, 50.75	< 0.001
C	44.50	42.89, 46.11	50.41	49.95, 50.88	< 0.001
Imp	54.25	52.56, 55.94	49.68	49.22, 50.15	< 0.001
SS	55.37	53.99, 56.74	49.60	49.13, 50.07	< 0.001
LSD					
#	69		1816		
N	50.28	47.82, 52.73	49.99	49.53, 50.45	0.820
E	52.70	49.90, 55.51	49.90	49.44, 50.35	0.053
O	57.56	55.78, 59.34	49.71	49.25, 50.17	< 0.001
A	50.03	47.88, 52.17	50.00	49.54, 50.46	0.979
C	46.98	44.90, 49.06	50.11	49.65, 50.58	0.004
Imp	52.51	50.18, 54.84	49.90	49.44, 50.36	0.032

Continued on the next page

Table C.4. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
SS	56.09	54.22, 57.96	49.77	49.31, 50.23	< 0.001
Methadone					
#	121		1764		
N	54.99	53.14, 56.84	49.66	49.20, 50.12	< 0.001
E	45.27	43.11, 47.43	50.32	49.87, 50.78	< 0.001
O	51.87	50.02, 53.72	49.87	49.41, 50.34	0.040
A	46.00	43.84, 48.16	50.27	49.82, 50.73	< 0.001
C	45.74	43.89, 47.59	50.29	49.83, 50.76	< 0.001
Imp	53.55	51.76, 55.34	49.76	49.29, 50.22	< 0.001
SS	53.93	52.18, 55.67	49.73	49.27, 50.20	< 0.001
Magic mushrooms					
#	44		1841		
N	49.79	46.82, 52.75	50.01	49.55, 50.46	0.884
E	53.71	50.31, 57.12	49.91	49.46, 50.37	0.031
O	57.89	55.73, 60.05	49.81	49.36, 50.27	< 0.001
A	50.14	46.71, 53.57	50.00	49.54, 50.45	0.935
C	48.03	45.33, 50.74	50.05	49.59, 50.51	0.146
Imp	55.44	52.56, 58.32	49.87	49.41, 50.33	< 0.001
SS	57.95	55.75, 60.15	49.81	49.35, 50.27	< 0.001
Nicotine					
#	767		1118		
N	51.32	50.59, 52.04	49.10	48.52, 49.67	< 0.001
E	49.91	49.17, 50.65	50.06	49.49, 50.63	0.748
O	51.57	50.85, 52.28	48.92	48.35, 49.50	< 0.001
A	49.04	48.31, 49.78	50.66	50.09, 51.23	0.001

Continued on the next page

Table C.4. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
C	47.69	46.99, 48.38	51.59	51.01, 52.16	< 0.001
Imp	52.56	51.84, 53.28	48.24	47.68, 48.80	< 0.001
SS	52.81	52.13, 53.50	48.07	47.50, 48.64	< 0.001
VSA					
#	21		1864		
N	50.92	46.02, 55.82	49.99	49.54, 50.44	0.697
E	52.59	47.66, 57.53	49.97	49.52, 50.42	0.283
O	56.30	53.63, 58.97	49.93	49.47, 50.38	< 0.001
A	46.36	41.38, 51.34	50.04	49.59, 50.49	0.141
C	49.99	45.49, 54.49	50.00	49.55, 50.45	0.995
Imp	55.21	51.37, 59.04	49.94	49.49, 50.40	0.010
SS	59.05	56.49, 61.61	49.90	49.44, 50.35	< 0.001
Heroin pleiad					
#	184		1701		
N	54.58	53.07, 56.08	49.50	49.04, 49.97	< 0.001
E	47.18	45.42, 48.95	50.30	49.84, 50.77	0.001
O	52.13	50.69, 53.57	49.77	49.29, 50.24	0.002
A	45.57	43.99, 47.16	50.48	50.01, 50.94	< 0.001
C	45.57	44.13, 47.01	50.48	50.01, 50.95	< 0.001
Imp	54.25	52.81, 55.68	49.54	49.07, 50.01	< 0.001
SS	55.04	53.66, 56.41	49.46	48.98, 49.93	< 0.001
Ecstasy pleiad					
#	792		1093		
N	50.94	50.22, 51.67	49.32	48.74, 49.89	0.001
E	49.69	48.94, 50.43	50.23	49.66, 50.79	0.255

Continued on the next page

Table C.4. *Continued*

Factor	Users		Non-users		*p*-value
	MT	95% CI	MT	95% CI	
O	54.07	53.46, 54.69	47.05	46.47, 47.63	< 0.001
A	48.34	47.63, 49.06	51.20	50.63, 51.77	< 0.001
C	47.08	46.38, 47.78	52.12	51.56, 52.67	< 0.001
Imp	53.23	52.55, 53.92	47.66	47.10, 48.22	< 0.001
SS	54.64	54.05, 55.24	46.63	46.06, 47.21	< 0.001
Benzodiazepines pleiad					
#	363		1522		
N	54.57	53.51, 55.63	48.91	48.43, 49.39	< 0.001
E	47.72	46.57, 48.87	50.54	50.06, 51.03	< 0.001
O	52.13	51.09, 53.16	49.49	48.99, 49.99	< 0.001
A	46.76	45.68, 47.84	50.77	50.28, 51.26	< 0.001
C	45.64	44.58, 46.69	51.04	50.55, 51.53	< 0.001
Imp	54.50	53.48, 55.52	48.93	48.44, 49.42	< 0.001
SS	54.24	53.29, 55.20	48.99	48.49, 49.49	< 0.001
Illicit drugs					
#	862		1023		
N	51.41	50.70, 52.12	48.76	48.19, 49.33	< 0.001
E	49.28	48.56, 50.01	51.13	50.59, 51.67	0.001
O	53.67	53.06, 54.27	46.07	45.47, 46.66	< 0.001
A	48.34	47.65, 49.03	51.19	50.60, 51.78	< 0.001
C	47.09	46.41, 47.77	53.07	52.53, 53.60	< 0.001
Imp	53.08	52.42, 53.73	46.72	46.16, 47.28	< 0.001
SS	54.35	53.76, 54.93	45.22	44.65, 45.79	< 0.001

C.2 Correlation between consumption of different drugs

In this section we show Pearson's correlation coefficients (PCCs) between drug consumptions for decade- and year-based user/non-user separation.

Table C.5. PCCs between drug consumptions with decade-based user/non-user separation. Amph. stays for amphetamines and Benz. for Benzodiazepines.

Drug	Amphetamines	Amyl nitrite	Benzodiazepines	Cannabis	Chocolate	Cocaine	Caffeine	Crack	Ecstasy	Heroin	Ketamine	Legal highs	LSD	Methadone	MMushrooms	Nicotine	VSA
Alcohol	0.074^{1}	0.074^{1}	0.051^{2}	0.119^{1}	0.099^{1}	0.111^{1}	0.157^{1}	0.027^{4}	0.105^{1}	0.033^{4}	0.078^{2}	0.061^{2}	0.069^{2}	-0.007^{4}	0.071^{1}	0.113^{1}	0.046^{3}
Amph.		0.372^{1}	0.463^{1}	0.469^{1}	0.013^{4}	0.580^{1}	0.106^{1}	0.323^{1}	0.597^{1}	0.359^{1}	0.412^{1}	0.481^{1}	0.490^{1}	0.415^{1}	0.481^{1}	0.343^{1}	0.304^{1}
Amyl nitrite			0.226^{1}	0.292^{1}	0.028^{4}	0.381^{1}	0.060^{2}	0.144^{1}	0.392^{1}	0.137^{1}	0.345^{1}	0.268^{1}	0.213^{1}	0.084^{1}	0.271^{1}	0.196^{1}	0.130^{1}
Benz.				0.354^{1}	0.006^{4}	0.428^{1}	0.055^{2}	0.326^{1}	0.383^{1}	0.395^{1}	0.303^{1}	0.348^{1}	0.352^{1}	0.468^{1}	0.366^{1}	0.260^{1}	0.294^{1}
Cannabis					0.046^{3}	0.453^{1}	0.113^{1}	0.216^{1}	0.521^{1}	0.217^{1}	0.302^{1}	0.526^{1}	0.421^{1}	0.299^{1}	0.497^{1}	0.533^{1}	0.237^{1}
Chocolate						0.006^{4}	0.122^{1}	0.032^{4}	0.040^{4}	-0.026^{4}	0.035^{4}	0.017^{4}	0.029^{4}	0.007^{4}	0.024^{4}	0.037^{4}	-0.021^{4}
Cocaine							0.099^{1}	0.396^{1}	0.633^{1}	0.414^{1}	0.454^{1}	0.445^{1}	0.442^{1}	0.354^{1}	0.480^{1}	0.362^{1}	0.277^{1}
Caffeine								0.035^{4}	0.107^{1}	0.026^{4}	0.058^{3}	0.085^{1}	0.075^{1}	0.039^{4}	0.100^{1}	0.145^{3}	0.053^{3}
Crack									0.280^{1}	0.509^{1}	0.255^{1}	0.203^{1}	0.268^{1}	0.367^{1}	0.276^{1}	0.191^{1}	0.278^{1}
Ecstasy										0.301^{1}	0.511^{1}	0.586^{1}	0.599^{1}	0.315^{1}	0.599^{1}	0.370^{1}	0.289^{1}
Heroin											0.274^{1}	0.237^{1}	0.347^{1}	0.494^{1}	0.306^{1}	0.185^{1}	0.293^{1}
Ketamine												0.393^{1}	0.462^{1}	0.246^{1}	0.436^{1}	0.243^{1}	0.192^{1}
Legal highs													0.519^{1}	0.334^{1}	0.575^{1}	0.364^{1}	0.314^{1}
LSD														0.343^{1}	0.680^{1}	0.289^{1}	0.299^{1}
Methadone															0.343^{1}	0.234^{1}	0.277^{1}
MMushrooms																0.324^{1}	0.253^{1}
Nicotine																	0.221^{1}

Note: $^{1}p < 0.001$, $^{2}p < 0.01$, $^{3}p < 0.05$, $^{4}p > 0.05$. p-value is the probability to observe by chance the same or greater correlation for uncorrelated variables.

Table C.6. PCCs between drug consumptions with year-based user/non-user separation. Amph. stays for amphetamines and Benz. for Benzodiazepines.

Drug	Amphetamines	Amyl nitrite	Benzodiazepines	Cannabis	Chocolate	Cocaine	Caffeine	Crack	Ecstasy	Heroin	Ketamine	Legal highs	LSD	Methadone	MMushrooms	Nicotine	VSA
Alcohol	0.046^{4}	0.061^{3}	0.048^{2}	0.078^{3}	0.077^{1}	0.124^{1}	0.111^{1}	0.058^{1}	0.107^{1}	0.030^{1}	0.072^{4}	0.093^{1}	0.084^{1}	-0.005^{1}	0.075^{4}	0.081^{1}	0.055^{3}
Amph.		0.222^{1}	0.436^{1}	0.421^{1}	0.003^{4}	0.453^{1}	0.072^{2}	0.193^{1}	0.461^{1}	0.305^{1}	0.325^{1}	0.471^{1}	0.392^{1}	0.382^{1}	0.375^{1}	0.311^{1}	0.173^{1}
Amyl nitrite			0.199^{1}	0.185^{1}	0.016^{4}	0.262^{1}	0.050^{3}	0.077^{1}	0.276^{1}	0.100^{1}	0.280^{1}	0.277^{1}	0.120^{1}	0.091^{1}	0.159^{1}	0.139^{1}	0.107^{1}
Benz,				0.334^{1}	-0.009^{4}	0.365^{1}	0.062^{2}	0.232^{1}	0.304^{1}	0.318^{1}	0.263^{1}	0.318^{1}	0.212^{1}	0.464^{1}	0.271^{1}	0.261^{1}	0.183^{1}
Cannabis					0.020^{4}	0.392^{1}	0.074^{2}	0.165^{1}	0.484^{1}	0.199^{1}	0.277^{1}	0.516^{1}	0.433^{1}	0.301^{1}	0.470^{1}	0.517^{1}	0.164^{1}
Chocolate						0.008^{4}	0.089^{1}	-0.037^{4}	0.057^{3}	-0.003^{4}	0.000^{4}	0.026^{4}	0.044^{4}	0.006^{4}	0.019^{4}	0.009^{4}	-0.012^{4}
Cocaine							0.069^{2}	0.322^{1}	0.535^{1}	0.358^{1}	0.379^{1}	0.394^{1}	0.302^{1}	0.314^{1}	0.346^{1}	0.329^{1}	0.187^{1}
Caffeine								0.023^{4}	0.059^{3}	0.023^{4}	0.026^{4}	0.067^{3}	0.032^{4}	0.027^{4}	0.050^{3}	0.105^{1}	0.042^{4}
Crack									0.156^{1}	0.350^{1}	0.180^{1}	0.147^{1}	0.139^{1}	0.265^{1}	0.181^{1}	0.126^{1}	0.145^{1}
Ecstasy										0.190^{1}	0.455^{1}	0.502^{1}	0.509^{1}	0.245^{1}	0.480^{1}	0.343^{1}	0.174^{1}
Heroin											0.217^{1}	0.185^{1}	0.170^{1}	0.385^{1}	0.171^{1}	0.149^{1}	0.121^{1}
Ketamine												0.373^{1}	0.351^{1}	0.202^{1}	0.362^{1}	0.222^{1}	0.151^{1}
Legal highs													0.434^{1}	0.309^{1}	0.485^{1}	0.348^{1}	0.220^{1}
LSD														0.234^{1}	0.627^{1}	0.267^{1}	0.174^{1}
Methadone															0.253^{1}	0.211^{1}	0.167^{1}
MMushrooms																0.282^{1}	0.174^{1}
Nicotine																	0.145^{1}

Note: $^{1}p < 0.001$, $^{2}p < 0.01$, $^{3}p < 0.05$, $^{4}p > 0.05$. p-value is the probability to observe by chance the same or greater correlation for uncorrelated variables.

C.3 Linear discriminants for user/non-user separation

Linear discriminants separate users from non-users by linear inequalities:

$$\Theta + \sum c_i z_i > 0$$

for users and ≤ 0 for non-users, where Θ are the threschholds, z_i are the attributes, and c_i are the coefficients. Tables C.7-C.10 contain the coefficients c_i of linear discriminants for user/nonuser separation in 10-dimensional space (7 psychological attributes, age, education, Gndr). The attributes in these tables are quantified and transformed to z-scores with zero mean and unite variance (positive values of the Gndr z-score corresponds to female). The last rows of the tables include the standard deviation of the coefficients in LOOCV. For 7-dimensional space of psychological attributes taken separately (T-scores), the linear discriminants are presented in tables C.15-C.18.

Performance of linear discriminants in user/non-user separation is evaluated by several methods (tables C.11–C.14 for 10-dimensional data space and tables C.19–C.22 for 7-dimensional space of T-scores of psychological attributes). First of all, we calculated the linear discriminant using the whole sample (see tables C.7–C.10) and find all their errors. For each solution of the classification problem we have several numbers, P (positive), the number of samples recognised as positive, and N - negative, the number of samples recognised as negative. $P + N$ is the total number of samples. P=TP+FP (True Positive plus False Positive) and N=TN+FN (True Negative plus False Negative). Sensitivity is Sn=TP/(TP+FN)$\times$100% and Specificity is Sp=TN/(TN+FP)$\times$100%. Accuracy is Acc=(TP+TN)/(P+N)$\times$100%. We calculate these performance indicators for the total sample and for the LOOCV procedure. In LOOCV the linear discriminant is calculated for the set of all samples excluding the example left out for testing. The test was performed for all samples with the corresponding redefining of Sn, Sp, and Acc. In LOOCV the linear discriminants are calculated for each testing example. Each of these dis-

criminants is a separate classification model. Stability of classification can be measured by the number of examples which change their class at least once. We took the basis model for the total sample and find how many true positive examples of this model became FN examples of a LOOCV model at least once. This number measured in % of TP+FP of the basic model is TP→FN. Analogously, we defined FP→TN, TN→FP, and FN→TP. The last two numbers are measured in % of TN+FN of the basic model.

Table C.7. Coefficients of linear discriminant for user/non-user separation and decade-based definition of users (10 attributes)

Drug	Θ	Age	Edu	N	E	O	A	C	Imp	SS	Gndr
Alcohol	0.358	-0.463	0.266	-0.042	0.030	-0.103	-0.117	-0.361	-0.031	0.710	0.230
Amphetamines	-0.236	-0.569	-0.236	0.081	-0.146	0.307	-0.017	-0.202	0.119	0.559	-0.362
Amyl nitrite	-0.261	-0.549	0.251	0.110	0.180	-0.142	-0.031	-0.229	-0.063	0.556	-0.451
Benz.	-0.098	-0.067	-0.243	0.566	-0.152	0.595	-0.164	-0.128	0.088	0.378	-0.214
Cannabis	0.434	-0.639	-0.241	-0.043	-0.213	0.518	-0.127	-0.193	0.043	0.352	-0.196
Chocolate	0.210	-0.169	-0.387	-0.149	-0.303	0.673	0.135	-0.168	-0.408	0.109	0.170
Cocaine	-0.198	-0.441	-0.121	0.220	0.068	0.276	-0.287	-0.235	0.090	0.682	-0.226
Caffeine	0.547	-0.522	-0.001	-0.262	0.388	0.331	0.070	-0.554	-0.074	0.266	0.097
Crack	-0.659	0.059	-0.577	0.175	-0.129	0.308	-0.088	-0.112	0.384	0.435	-0.405
Ecstasy	-0.171	-0.631	-0.188	0.053	0.018	0.351	-0.065	-0.210	-0.088	0.559	-0.265
Heroin	-0.615	-0.210	-0.370	0.413	-0.211	0.477	-0.265	-0.029	0.222	0.381	-0.332
Ketamine	-0.448	-0.560	-0.005	0.065	-0.022	0.367	-0.139	-0.223	0.037	0.507	-0.469
Legal highs	-0.172	-0.655	-0.131	0.015	-0.174	0.427	0.006	-0.143	-0.052	0.429	-0.366
LSD	-0.409	-0.581	-0.168	-0.069	-0.203	0.539	-0.060	-0.027	-0.000	0.419	-0.344
Methadone	-0.433	-0.346	-0.382	0.225	-0.336	0.564	-0.212	-0.118	0.088	0.274	-0.332
M. mushrooms	-0.243	-0.616	-0.134	-0.119	-0.186	0.541	-0.064	-0.139	0.075	0.358	-0.321
Nicotine	0.324	-0.657	-0.365	0.123	-0.072	0.364	-0.082	-0.196	0.026	0.450	-0.184
VSA	-0.818	-0.805	-0.143	0.059	-0.145	0.207	-0.066	-0.189	0.023	0.441	-0.172
Heroin pleiad	-0.082	-0.377	-0.240	0.146	-0.117	0.450	-0.260	-0.240	0.062	0.605	-0.261
Ecstasy pleiad	0.494	-0.632	-0.223	-0.041	-0.201	0.504	-0.164	-0.228	0.034	0.372	-0.191
Benz. pleiad	0.151	-0.316	-0.292	0.274	-0.100	0.524	-0.175	-0.233	-0.008	0.552	-0.257
Illicit drugs	0.624	-0.573	-0.263	0.069	-0.243	0.523	-0.142	-0.243	0.070	0.360	-0.228
SD	≤ 0.009	≤ 0.006	≤ 0.006	≤ 0.007	≤ 0.006	≤ 0.005	≤ 0.006	≤ 0.007	≤ 0.007	≤ 0.008	≤ 0.006

Table C.8. Coefficients of linear discriminant for user/non-user separation and year-based definition of users (10 attributes)

Drug	Θ	Age	Edu	N	E	O	A	C	Imp	SS	Gndr
Alcohol	0.218	-0.585	0.342	-0.228	0.101	-0.133	-0.031	-0.165	-0.211	0.575	0.241
Amphetamines	-0.543	-0.643	-0.249	0.063	-0.176	0.347	-0.103	-0.201	0.241	0.418	-0.293
Amyl nitrite	-0.527	-0.647	0.085	0.124	0.075	-0.036	-0.249	-0.142	-0.191	0.420	-0.507
Benz.	-0.296	-0.272	-0.263	0.594	-0.110	0.437	-0.187	-0.007	0.102	0.419	-0.282
Cannabis	0.130	-0.569	-0.310	-0.069	-0.190	0.521	-0.020	-0.194	-0.034	0.411	-0.245
Chocolate	0.169	-0.259	-0.217	0.036	-0.168	0.575	0.064	-0.236	-0.571	0.072	0.367
Cocaine	-0.494	-0.685	-0.007	0.211	0.177	0.077	-0.247	-0.175	0.086	0.521	-0.280
Caffeine	0.517	-0.323	0.129	-0.305	0.469	0.039	-0.061	-0.602	0.079	0.434	0.063
Crack	-0.973	-0.153	-0.408	0.257	-0.014	0.090	-0.060	-0.079	0.346	0.376	-0.682
Ecstasy	-0.464	-0.782	-0.101	-0.015	0.099	0.238	-0.025	-0.173	-0.004	0.453	-0.275
Heroin	-0.849	-0.584	-0.168	0.352	-0.252	0.222	-0.275	0.014	0.216	0.359	-0.378
Ketamine	-0.675	-0.790	0.033	0.070	0.046	0.214	-0.075	-0.189	0.038	0.354	-0.392
Legal highs	-0.432	-0.656	-0.212	-0.035	-0.134	0.370	-0.054	-0.129	-0.026	0.481	-0.342
LSD	-0.685	-0.757	-0.128	-0.112	-0.137	0.451	0.001	0.040	-0.015	0.302	-0.288
Methadone	-0.527	-0.450	-0.322	0.244	-0.331	0.495	-0.246	-0.051	0.050	0.289	-0.360
M. mushrooms	-0.573	-0.712	-0.197	-0.106	-0.119	0.490	-0.033	-0.053	0.040	0.293	-0.310
Nicotine	0.149	-0.579	-0.446	0.096	-0.076	0.345	-0.017	-0.255	0.011	0.472	-0.209
VSA	-0.847	-0.859	-0.111	0.268	0.002	0.087	-0.032	-0.095	-0.016	0.315	-0.247
Heroin pleiad	-0.368	-0.601	-0.190	0.176	-0.088	0.299	-0.267	-0.169	0.092	0.506	-0.333
Ecstasy pleiad	0.230	-0.576	-0.296	-0.062	-0.222	0.512	-0.077	-0.185	0.013	0.396	-0.261
Benz. pleiad	-0.069	-0.449	-0.284	0.292	-0.119	0.429	-0.207	-0.156	0.073	0.537	-0.271
Illicit drugs	0.348	-0.524	-0.335	0.032	-0.248	0.528	-0.092	-0.182	0.036	0.395	-0.271
SD	≤ 0.014	≤ 0.007	≤ 0.007	≤ 0.009	≤ 0.008	≤ 0.007	≤ 0.008	≤ 0.008	≤ 0.007	≤ 0.010	≤ 0.007

Table C.9. Coefficients of linear discriminant for user/non-user separation and month-based definition of users (10 attributes)

Drug	Θ	Age	Edu	N	E	O	A	C	Imp	SS	Gndr
Alcohol	0.130	-0.263	0.590	0.096	0.588	-0.111	0.078	-0.083	-0.193	0.402	0.058
Amphetamines	-0.543	-0.643	-0.249	0.063	-0.176	0.347	-0.103	-0.201	0.241	0.418	-0.293
Amyl nitrite	-0.821	-0.361	-0.229	-0.223	-0.114	-0.178	-0.144	-0.018	-0.088	0.749	-0.365
Benz.	-0.416	-0.115	-0.243	0.711	-0.128	0.284	-0.180	0.072	0.167	0.418	-0.292
Cannabis	-0.122	-0.542	-0.394	-0.132	-0.166	0.547	-0.037	-0.132	0.015	0.355	-0.250
Chocolate	0.132	0.138	-0.284	-0.161	0.107	0.379	-0.004	-0.440	-0.488	0.193	0.501
Cocaine	-0.597	-0.624	0.029	0.345	0.212	-0.007	-0.305	0.054	0.062	0.523	-0.270
Caffeine	0.273	-0.019	0.369	0.261	0.637	-0.043	0.035	-0.239	0.424	0.385	0.042
Crack	-0.836	0.154	-0.131	0.449	-0.114	-0.075	-0.253	0.156	0.076	0.581	-0.555
Ecstasy	-0.633	-0.820	0.047	-0.139	0.093	0.284	-0.123	-0.165	-0.028	0.328	-0.257
Heroin	-1.037	-0.560	-0.371	0.181	-0.350	0.159	-0.397	0.016	0.368	0.154	-0.226
Ketamine	-0.793	-0.776	0.020	-0.097	-0.147	0.340	-0.098	-0.268	-0.039	0.139	-0.386
Legal highs	-0.693	-0.519	-0.224	-0.012	-0.190	0.409	-0.136	-0.240	0.022	0.427	-0.467
LSD	-0.851	-0.722	-0.173	0.006	-0.045	0.541	0.007	-0.032	-0.098	0.252	-0.284
Methadone	-0.551	-0.404	-0.259	0.262	-0.443	0.417	-0.270	-0.105	0.003	0.399	-0.296
M. mushrooms	-0.764	-0.594	-0.233	-0.184	-0.236	0.604	-0.019	-0.066	0.070	0.239	-0.267
Nicotine	-0.019	-0.461	-0.530	0.092	0.037	0.318	0.013	-0.375	0.157	0.389	-0.283
VSA	-1.027	-0.785	0.003	0.081	0.059	0.174	-0.215	0.073	0.096	0.482	-0.222
Heroin pleiad	-0.545	-0.525	-0.187	0.301	-0.180	0.224	-0.295	0.013	0.129	0.530	-0.364
Ecstasy pleiad	0.019	-0.576	-0.339	-0.133	-0.207	0.514	-0.098	-0.172	0.073	0.355	-0.241
Benz. pleiad	-0.346	-0.309	-0.254	0.479	-0.125	0.274	-0.215	-0.123	0.171	0.534	-0.380
Illicit drugs	0.111	-0.513	-0.390	-0.034	-0.236	0.532	-0.097	-0.151	0.061	0.371	-0.267
SD	≤ 0.008	≤ 0.006	≤ 0.005	≤ 0.006	≤ 0.006	≤ 0.006	≤ 0.006	≤ 0.006	≤ 0.008	≤ 0.008	≤ 0.006

Table C.10. Coefficients of linear discriminant for user/non-user separation and week-based definition of users (10 attributes)

Drug	Θ	Age	Edu	N	E	O	A	C	Imp	SS	Gndr
Alcohol	0.063	0.053	0.640	0.157	0.522	-0.283	0.069	-0.125	-0.105	0.412	-0.099
Amphetamines	-0.546	-0.441	-0.070	-0.071	-0.295	0.348	-0.207	-0.359	0.584	0.120	-0.248
Amyl nitrite	-0.747	0.023	0.173	-0.389	-0.351	-0.203	0.006	-0.213	0.214	0.695	-0.281
Benz.	-0.546	0.027	-0.331	0.680	-0.371	0.273	-0.230	0.091	0.231	0.287	-0.137
Cannabis	-0.240	-0.502	-0.402	-0.120	-0.172	0.594	-0.036	-0.108	-0.023	0.335	-0.253
Chocolate	0.109	0.390	0.195	0.117	0.444	-0.034	0.114	-0.172	0.283	-0.461	0.513
Cocaine	-0.806	-0.543	0.015	0.358	0.274	-0.118	-0.299	-0.048	-0.042	0.534	-0.331
Caffeine	0.131	0.484	0.200	0.335	0.637	-0.079	-0.166	-0.277	0.088	0.273	0.122
Crack	-1.170	0.151	-0.096	0.258	-0.068	-0.347	-0.312	0.133	0.111	0.438	-0.676
Ecstasy	-0.779	-0.697	0.077	-0.115	0.161	0.545	-0.093	-0.252	0.217	0.230	-0.022
Heroin	-1.096	-0.386	-0.077	0.467	-0.255	0.184	-0.412	0.013	0.437	-0.077	-0.400
Ketamine	-0.732	-0.665	0.074	-0.176	-0.223	0.380	-0.264	-0.239	-0.087	0.117	-0.423
Legal highs	-0.755	-0.377	-0.252	0.037	-0.302	0.300	-0.253	-0.295	-0.016	0.422	-0.531
LSD	-1.085	-0.679	-0.310	0.039	0.043	0.421	0.053	-0.054	-0.168	0.127	-0.462
Methadone	-0.606	-0.359	-0.247	0.290	-0.549	0.380	-0.242	-0.024	0.085	0.257	-0.384
M. mushrooms	-0.657	-0.296	-0.281	-0.136	-0.046	0.705	0.036	-0.030	0.070	0.529	-0.167
Nicotine	-0.040	-0.252	-0.653	0.201	0.075	0.233	0.052	-0.335	0.194	0.403	-0.308
VSA	-1.047	-0.620	-0.042	0.076	0.016	0.270	-0.204	0.301	-0.021	0.555	-0.306
Heroin pleiad	-0.680	-0.390	-0.234	0.360	-0.304	0.248	-0.309	-0.119	0.136	0.431	-0.442
Ecstasy pleiad	-0.113	-0.524	-0.387	-0.127	-0.219	0.545	-0.093	-0.159	0.047	0.341	-0.251
Benz. pleiad	-0.423	-0.269	-0.305	0.504	-0.294	0.260	-0.233	-0.163	0.381	0.362	-0.265
Illicit drugs	-0.051	-0.471	-0.414	-0.029	-0.266	0.540	-0.080	-0.142	0.048	0.378	-0.268
SD	≤ 0.018	≤ 0.006	≤ 0.005	≤ 0.005	≤ 0.005	≤ 0.005	≤ 0.006	≤ 0.005	≤ 0.006	≤ 0.006	≤ 0.005

Table C.11. Performance and stability of linear discriminant for decade-based definition of users (10 attributes).

	Total sample			LOOCV			Stability indicators			
Drug	Sn	Sp	Acc	Sn	Sp	Acc	TP→FN	FN→TP	FP→TN	TN→FP
Alcohol	65.1	67.6	65.2	64.9	57.4	64.6	5.8	8.5	11.8	7.4
Amphetamines	72.0	71.9	71.9	71.1	71.5	71.4	1.2	0.9	0.7	0.5
Amyl nitrite	63.8	63.6	63.6	62.2	63.0	62.9	2.4	1.6	1.3	1.8
Benz.	67.4	67.5	67.4	66.8	67.1	67.0	0.4	0.4	0.6	1.3
Cannabis	78.1	78.2	78.1	77.9	77.6	77.8	0.4	0.6	0.8	1.0
Chocolate	57.9	57.1	57.9	57.5	37.1	57.1	11.2	14.3	14.1	28.6
Cocaine	67.7	67.6	67.6	67.0	67.3	67.2	0.9	0.6	0.5	1.0
Caffeine	68.7	70.3	68.8	68.6	62.2	68.5	5.5	8.8	8.1	13.5
Crack	69.6	69.1	69.1	65.4	68.9	68.5	5.2	2.1	1.7	1.7
Ecstasy	74.4	74.7	74.6	74.2	74.3	74.2	0.3	0.7	0.5	0.4
Heroin	70.8	69.9	70.0	68.4	69.8	69.7	2.8	6.1	1.3	2.2
Ketamine	68.0	67.8	67.8	66.6	67.5	67.3	1.4	1.4	0.5	1.4
Legal highs	79.0	79.2	79.1	78.9	78.9	78.9	0.1	0.8	0.3	0.4
LSD	76.8	76.7	76.8	76.5	76.7	76.6	0.5	0.7	0.5	0.7
Methadone	70.5	70.8	70.7	69.3	70.2	70.0	1.2	0.7	1.4	0.5
M. mushrooms	75.6	75.7	75.6	75.1	75.3	75.2	0.9	0.1	0.7	0.5
Nicotine	70.1	70.4	70.2	69.7	69.1	69.5	0.8	0.9	1.6	1.1
VSA	74.8	74.8	74.8	74.3	74.7	74.5	2.2	2.2	0.7	1.6
Heroin pleiad	69.6	70.1	69.9	69.1	69.4	69.3	0.5	0.7	0.9	0.4
Ecstasy pleiad	78.6	78.7	78.6	78.6	78.2	78.5	0.5	0.8	0.4	0.4
Benz. pleiad	70.7	70.7	70.7	70.2	69.8	70.0	0.9	0.8	0.9	0.8
Illicit drugs	77.9	78.2	78.1	77.5	77.8	77.7	0.4	0.6	0.7	0.9

Table C.12. Performance and stability of linear discriminant for year-based definition of users (10 attributes).

	Total sample			LOOCV			Stability indicators			
Drug	Sn	Sp	Acc	Sn	Sp	Acc	TP→FN	FN→TP	FP→TN	TN→FP
Alcohol	63.1	63.2	63.1	62.7	58.1	62.4	3.3	4.6	5.9	4.4
Amphetamines	72.5	72.7	72.7	71.6	72.4	72.2	0.9	0.5	0.9	0.9
Amyl nitrite	69.2	68.6	68.6	62.4	68.3	67.9	4.5	2.3	2.7	2.6
Benz.	69.7	69.4	69.5	68.2	69.3	69.0	1.9	0.6	0.5	0.7
Cannabis	79.3	79.3	79.3	78.9	79.0	78.9	0.7	0.8	0.3	0.2
Chocolate	58.1	60.0	58.1	57.4	37.8	57.0	13.6	16.1	17.8	13.3
Cocaine	71.0	71.5	71.4	69.8	71.1	70.8	1.0	0.2	1.0	1.2
Caffeine	68.0	67.2	68.0	67.7	54.1	67.2	8.9	7.3	11.5	6.6
Crack	73.4	73.9	73.9	69.6	73.6	73.4	1.3	2.5	3.4	3.7
Ecstasy	75.4	75.7	75.6	75.0	75.6	75.4	0.2	0.6	0.8	0.3
Heroin	73.7	73.7	73.7	70.3	73.6	73.4	2.5	2.5	2.1	2.3
Ketamine	71.6	71.6	71.6	68.8	71.4	71.1	2.9	1.9	2.1	1.9
Legal highs	77.7	77.7	77.7	77.0	77.4	77.3	0.9	0.4	0.5	0.5
LSD	80.0	80.0	80.0	79.2	79.9	79.7	0.5	0.5	0.3	0.9
Methadone	70.6	70.6	70.6	69.7	70.4	70.3	0.9	0.9	0.7	1.7
M. mushrooms	77.9	77.9	77.9	77.4	77.7	77.7	0.5	0.2	0.2	0.6
Nicotine	70.4	70.4	70.4	70.1	70.2	70.1	0.3	0.8	0.4	0.4
VSA	72.6	72.7	72.7	71.6	72.5	72.4	2.1	2.1	1.7	3.9
Heroin pleiad	71.5	72.0	71.8	71.3	71.7	71.6	0.3	0.7	0.8	1.0
Ecstasy pleiad	80.2	80.3	80.2	80.0	80.0	80.0	0.1	0.6	0.3	0.5
Benz. pleiad	73.0	72.7	72.8	72.4	72.3	72.4	0.6	0.4	0.5	0.4
Illicit drugs	80.2	79.6	79.8	79.5	79.4	79.4	1.0	0.4	0.4	0.3

Table C.13. Performance and stability of linear discriminant for month-based definition of users (10 attributes).

	Total sample			LOOCV			Stability indicators			
Drug	Sn	Sp	Acc	Sn	Sp	Acc	TP→FN	FN→TP	FP→TN	TN→FP
Alcohol	60.9	61.1	60.9	60.4	58.4	60.1	2.3	2.6	3.0	2.4
Amphetamines	68.9	68.1	68.2	65.5	68.0	67.7	2.9	3.4	1.2	2.0
Amyl nitrite	70.7	75.3	75.2	63.4	75.1	74.9	2.4	2.4	8.5	5.2
Benz.	68.9	68.7	68.8	68.2	68.5	68.5	1.3	1.3	1.1	2.0
Cannabis	78.6	78.3	78.4	77.7	77.9	77.8	1.1	0.0	0.6	0.9
Chocolate	57.9	56.6	57.8	57.2	48.5	56.7	9.7	8.0	9.1	8.1
Cocaine	69.8	70.0	70.0	67.3	69.9	69.7	3.1	2.5	2.3	2.7
Caffeine	60.8	60.3	60.7	60.5	54.5	60.2	7.3	5.2	6.6	5.8
Crack	80.0	75.3	75.4	65.0	75.2	75.1	15.0	15.0	4.8	11.6
Ecstasy	72.5	72.4	72.4	71.3	72.3	72.1	1.7	1.3	1.2	1.2
Heroin	79.2	77.6	77.7	69.8	77.5	77.2	9.4	7.5	3.2	3.8
Ketamine	72.2	73.3	73.2	64.6	73.1	72.8	3.8	1.3	3.4	2.7
Legal highs	72.6	72.5	72.5	71.4	72.4	72.3	1.2	1.7	0.9	1.2
LSD	76.5	76.4	76.4	74.7	76.3	76.2	1.2	1.2	1.0	1.9
Methadone	68.4	68.1	68.2	66.7	67.7	67.6	1.8	3.5	1.6	2.6
M. mushrooms	75.5	74.3	74.4	73.6	74.3	74.2	2.5	2.5	0.9	2.0
Nicotine	66.4	66.2	66.3	65.8	66.0	65.9	1.1	0.7	0.3	0.7
VSA	76.5	75.6	75.6	58.8	75.6	75.3	5.9	5.9	2.9	6.1
Heroin pleiad	69.3	69.4	69.4	68.0	69.2	69.0	1.6	1.3	1.5	0.9
Ecstasy pleiad	79.4	79.6	79.5	79.2	79.1	79.2	0.5	0.5	0.7	0.7
Benz. pleiad	70.5	70.3	70.3	70.1	69.9	70.0	0.4	0.8	0.7	1.0
Illicit drugs	78.9	78.8	78.8	78.4	78.5	78.5	0.4	0.3	0.5	0.4

Table C.14. Performance and stability of linear discriminant for week-based definition of users (10 attributes).

	Total sample			LOOCV			Stability indicators			
Drug	Sn	Sp	Acc	Sn	Sp	Acc	TP→FN	FN→TP	FP→TN	TN→FP
Alcohol	60.4	59.9	60.3	59.3	58.9	59.2	2.0	0.9	1.0	1.3
Amphetamines	66.9	66.7	66.7	62.0	66.5	66.1	4.3	2.5	3.0	2.8
Amyl nitrite	70.6	77.6	77.5	52.9	77.4	77.1	0.0	5.9	14.7	11.8
Benz.	70.4	70.1	70.1	67.0	69.8	69.5	3.4	1.7	2.1	2.7
Cannabis	75.9	75.9	75.9	75.3	75.5	75.4	0.9	0.0	0.5	0.5
Chocolate	57.8	57.7	57.8	57.2	54.4	56.6	3.6	2.5	3.3	1.3
Cocaine	75.0	75.1	75.1	61.7	75.1	74.6	8.3	6.7	4.1	4.4
Caffeine	59.5	58.1	59.4	58.5	52.0	57.7	5.8	2.8	6.2	5.3
Crack	90.9	85.9	85.9	54.5	85.8	85.6	27.3	9.1	4.1	7.7
Ecstasy	72.6	71.5	71.5	67.9	71.4	71.2	3.6	8.3	2.4	5.3
Heroin	79.3	80.1	80.1	65.5	80.0	79.8	6.9	6.9	4.6	4.8
Ketamine	73.0	71.9	71.9	62.2	71.6	71.5	8.1	8.1	3.5	6.5
Legal highs	70.2	71.0	71.0	66.4	70.9	70.6	3.8	5.3	2.5	1.9
LSD	81.2	80.6	80.6	76.8	80.3	80.2	4.3	2.9	1.5	2.2
Methadone	70.2	69.4	69.5	65.3	69.3	69.1	3.3	4.1	2.6	3.2
M. mushrooms	70.5	68.7	68.7	61.4	68.4	68.2	6.8	9.1	4.4	7.9
Nicotine	63.9	63.8	63.8	63.5	63.5	63.5	0.5	0.8	0.3	0.6
VSA	81.0	78.7	78.7	61.9	78.6	78.4	4.8	0.0	5.2	7.7
Heroin pleiad	70.7	70.3	70.3	65.8	69.8	69.4	4.9	1.6	2.1	1.6
Ecstasy pleiad	77.3	77.3	77.3	76.6	77.0	76.9	0.6	0.3	0.6	0.8
Benz. pleiad	68.3	68.5	68.4	66.9	68.3	68.0	1.1	1.1	1.2	0.9
Illicit drugs	77.2	77.1	77.2	76.8	76.7	76.8	0.7	0.8	0.7	0.5

Table C.15. Coefficients of linear discriminant for user/non-user separation and decade-based definition of users (7 attributes)

Drug	Θ	N	E	O	A	C	Imp	SS
Alcohol	-22.282	0.124	0.149	-0.068	-0.102	-0.460	0.056	0.856
Amphetamines	-36.596	0.040	-0.197	0.363	-0.063	-0.343	0.040	0.840
Amyl nitrite	-29.255	0.083	0.177	-0.023	-0.104	-0.308	-0.195	0.904
Benz.	-56.705	0.532	-0.211	0.597	-0.205	-0.186	0.105	0.478
Cannabis	-22.278	-0.004	-0.242	0.571	-0.143	-0.357	-0.014	0.684
Chocolate	6.158	-0.133	-0.369	0.603	0.271	-0.308	-0.467	0.309
Cocaine	-43.257	0.203	0.037	0.303	-0.288	-0.293	0.024	0.834
Caffeine	-9.371	-0.121	0.479	0.313	0.026	-0.682	-0.139	0.416
Crack	-48.494	0.145	-0.293	0.326	-0.202	-0.243	0.415	0.718
Ecstasy	-35.896	0.045	-0.017	0.407	-0.085	-0.342	-0.156	0.827
Heroin	-55.851	0.381	-0.283	0.526	-0.322	-0.124	0.249	0.563
Ketamine	-34.711	-0.003	-0.094	0.478	-0.197	-0.333	-0.024	0.782
Legal highs	-30.803	-0.017	-0.227	0.532	-0.040	-0.303	-0.118	0.747
LSD	-34.972	-0.108	-0.258	0.632	-0.103	-0.162	-0.068	0.694
Methadone	-27.409	0.208	-0.416	0.620	-0.262	-0.253	0.058	0.513
M. mushrooms	-29.576	-0.147	-0.239	0.631	-0.106	-0.265	0.002	0.665
Nicotine	-38.644	0.171	-0.092	0.397	-0.093	-0.393	0.005	0.801
VSA	-34.395	0.071	-0.209	0.326	-0.060	-0.392	0.012	0.829
Heroin pleiad	-30.383	0.113	-0.149	0.441	-0.270	-0.331	0.007	0.767
Ecstasy pleiad	-17.510	-0.002	-0.222	0.542	-0.180	-0.385	-0.028	0.689
Benz. pleiad	-35.808	0.232	-0.136	0.492	-0.201	-0.347	-0.042	0.723
Illicit drugs	-18.501	0.096	-0.272	0.537	-0.169	-0.409	0.014	0.657
SD	≤ 0.970	≤ 0.008	≤ 0.008	≤ 0.007	≤ 0.007	≤ 0.008	≤ 0.008	≤ 0.009

Table C.16. Coefficients of linear discriminant for user/non-user separation and year-based definition of users (7 attributes)

Drug	Θ	N	E	O	A	C	Imp	SS
Alcohol	-16.181	-0.136	0.270	-0.082	-0.024	-0.236	-0.314	0.865
Amphetamines	-39.624	0.061	-0.213	0.424	-0.136	-0.365	0.166	0.769
Amyl nitrite	-5.472	0.060	0.023	0.157	-0.388	-0.292	-0.319	0.796
Benz.	-64.199	0.566	-0.171	0.456	-0.220	-0.112	0.089	0.611
Cannabis	-24.008	-0.064	-0.230	0.555	-0.040	-0.354	-0.093	0.706
Chocolate	0.709	0.114	-0.159	0.582	0.264	-0.380	-0.606	0.203
Cocaine	-46.414	0.224	0.152	0.187	-0.299	-0.292	0.030	0.846
Caffeine	-8.114	-0.222	0.503	0.076	-0.063	-0.636	0.052	0.530
Crack	-52.425	0.244	-0.153	0.140	-0.219	-0.282	0.399	0.782
Ecstasy	-42.579	-0.019	0.078	0.373	-0.045	-0.356	-0.096	0.846
Heroin	-44.733	0.361	-0.327	0.368	-0.397	-0.108	0.218	0.641
Ketamine	-42.342	0.056	-0.003	0.417	-0.143	-0.363	-0.047	0.818
Legal highs	-30.027	-0.062	-0.181	0.460	-0.091	-0.300	-0.098	0.803
LSD	-40.183	-0.141	-0.220	0.650	-0.039	-0.118	-0.115	0.693
Methadone	-28.093	0.234	-0.417	0.575	-0.296	-0.206	0.024	0.556
M. mushrooms	-36.864	-0.142	-0.182	0.669	-0.093	-0.215	-0.053	0.664
Nicotine	-35.594	0.122	-0.126	0.376	-0.036	-0.441	0.019	0.795
VSA	-55.631	0.382	0.009	0.213	-0.038	-0.379	-0.014	0.815
Heroin pleiad	-32.917	0.154	-0.147	0.385	-0.310	-0.312	0.044	0.782
Ecstasy pleiad	-20.439	-0.059	-0.264	0.548	-0.102	-0.354	-0.045	0.699
Benz. pleiad	-41.126	0.256	-0.164	0.438	-0.230	-0.287	0.018	0.761
Illicit drugs	-22.701	0.033	-0.295	0.549	-0.125	-0.351	-0.012	0.686
SD	≤ 1.250	≤ 0.011	≤ 0.011	≤ 0.009	≤ 0.009	≤ 0.009	≤ 0.008	≤ 0.012

Table C.17. Coefficients of linear discriminant for user/non-user separation and month-based definition of users (7 attributes)

Drug	Θ	N	E	O	A	C	Imp	SS
Alcohol	-63.256	0.190	0.823	0.002	0.087	-0.005	-0.268	0.456
Amphetamines	-24.230	0.011	-0.223	0.361	-0.254	-0.499	0.375	0.605
Amyl nitrite	-6.284	-0.243	-0.183	-0.082	-0.187	-0.092	-0.117	0.918
Benz.	-69.623	0.702	-0.189	0.286	-0.215	0.006	0.148	0.568
Cannabis	-25.637	-0.145	-0.215	0.606	-0.069	-0.322	-0.051	0.674
Chocolate	19.518	-0.121	0.254	0.308	0.168	-0.595	-0.621	0.240
Cocaine	-57.737	0.366	0.148	0.082	-0.370	-0.014	-0.001	0.837
Caffeine	-91.137	0.351	0.687	0.026	0.052	-0.080	0.567	0.271
Crack	-46.531	0.459	-0.220	-0.077	-0.357	0.096	0.066	0.771
Ecstasy	-27.572	-0.169	0.101	0.462	-0.191	-0.354	-0.139	0.752
Heroin	-30.020	0.302	-0.409	0.232	-0.499	-0.107	0.345	0.555
Ketamine	-6.182	-0.169	-0.210	0.643	-0.232	-0.464	-0.044	0.493
Legal highs	-17.306	-0.067	-0.240	0.474	-0.201	-0.388	-0.053	0.721
LSD	-47.031	-0.012	-0.097	0.762	-0.023	-0.225	-0.155	0.578
Methadone	-17.825	0.260	-0.501	0.436	-0.286	-0.238	-0.004	0.594
M. mushrooms	-29.897	-0.229	-0.273	0.762	-0.037	-0.217	-0.019	0.494
Nicotine	-34.684	0.072	-0.032	0.342	-0.022	-0.590	0.204	0.698
VSA	-70.559	0.137	0.068	0.230	-0.208	-0.073	0.182	0.918
Heroin pleiad	-41.444	0.277	-0.272	0.284	-0.346	-0.105	0.113	0.791
Ecstasy pleiad	-19.318	-0.141	-0.257	0.573	-0.125	-0.359	0.014	0.664
Benz. pleiad	-47.877	0.416	-0.199	0.300	-0.258	-0.246	0.128	0.745
Illicit drugs	-21.767	-0.051	-0.295	0.567	-0.129	-0.344	0.009	0.674
SD	≤ 1.041	≤ 0.009	≤ 0.009	≤ 0.009	≤ 0.008	≤ 0.007	≤ 0.007	≤ 0.010

Table C.18. Coefficients of linear discriminant for user/non-user separation and week-based definition of users (7 attributes)

Drug	Θ	N	E	O	A	C	Imp	SS
Alcohol	-52.758	0.202	0.791	-0.273	0.041	-0.002	-0.176	0.476
Amphetamines	-17.930	-0.054	-0.311	0.402	-0.237	-0.472	0.577	0.356
Amyl nitrite	5.116	-0.466	-0.348	-0.188	-0.084	-0.186	0.249	0.723
Benz.	-51.825	0.707	-0.438	0.240	-0.229	0.052	0.224	0.382
Cannabis	-28.488	-0.136	-0.215	0.658	-0.074	-0.282	-0.079	0.641
Chocolate	-22.538	0.221	0.565	-0.149	0.195	0.020	0.302	-0.693
Cocaine	-42.900	0.347	0.198	-0.011	-0.407	-0.090	-0.103	0.810
Caffeine	-53.485	0.422	0.821	-0.118	-0.135	-0.146	0.301	-0.061
Crack	-22.095	0.304	-0.279	-0.282	-0.463	0.148	0.119	0.707
Ecstasy	-60.127	-0.095	0.168	0.695	-0.128	-0.355	0.242	0.528
Heroin	-37.099	0.548	-0.325	0.265	-0.547	-0.069	0.389	0.261
Ketamine	1.992	-0.284	-0.395	0.574	-0.307	-0.228	-0.030	0.535
Legal highs	0.347	-0.015	-0.352	0.406	-0.315	-0.454	-0.022	0.636
LSD	-35.647	0.001	0.054	0.716	0.008	-0.371	-0.309	0.501
Methadone	-14.932	0.317	-0.636	0.331	-0.310	-0.106	0.084	0.521
M. mushrooms	-66.891	-0.143	-0.071	0.715	0.073	-0.114	0.059	0.665
Nicotine	-42.383	0.175	-0.013	0.222	0.012	-0.574	0.284	0.714
VSA	-75.762	0.043	-0.068	0.419	-0.191	0.250	0.056	0.846
Heroin pleiad	-25.458	0.341	-0.397	0.242	-0.379	-0.205	0.116	0.684
Ecstasy pleiad	-19.877	-0.146	-0.272	0.603	-0.124	-0.338	-0.004	0.642
Benz. pleiad	-43.258	0.498	-0.349	0.260	-0.250	-0.274	0.343	0.554
Illicit drugs	-22.772	-0.055	-0.329	0.575	-0.113	-0.322	0.004	0.665
SD	≤ 0.707	≤ 0.006	≤ 0.006	≤ 0.006	≤ 0.007	≤ 0.007	≤ 0.009	≤ 0.007

Table C.19. Performance and stability of linear discriminant for decade-based definition of users (7 attributes).

	Total sample			LOOCV			Stability indicators			
Drug	Sn	Sp	Acc	Sn	Sp	Acc	TP→FN	FN→TP	FP→TN	TN→FP
Alcohol	67.1	64.7	67.0	66.9	55.9	66.5	7.6	4.7	11.8	4.4
Amphetamines	69.1	69.3	69.2	68.5	68.7	68.6	0.7	1.0	1.1	1.2
Amyl nitrite	61.9	61.8	61.9	61.1	61.7	61.5	1.6	2.4	0.9	2.0
Benz.	67.6	67.3	67.4	67.0	66.8	66.9	0.8	0.7	0.8	0.9
Cannabis	75.0	75.2	75.1	74.9	74.4	74.7	0.5	0.5	0.8	0.6
Chocolate	56.7	54.3	56.7	56.4	42.9	56.2	11.4	13.1	5.7	20.0
Cocaine	66.4	66.5	66.5	66.1	66.2	66.2	0.4	0.6	0.7	0.7
Caffeine	69.9	70.3	69.9	69.7	56.8	69.4	6.5	7.5	10.8	8.1
Crack	66.0	65.9	65.9	62.8	65.6	65.3	4.7	1.6	2.1	2.6
Ecstasy	71.6	71.6	71.6	71.4	71.4	71.4	0.5	0.8	0.4	0.7
Heroin	70.3	70.8	70.7	68.4	70.6	70.3	1.9	1.4	2.1	1.6
Ketamine	66.9	66.8	66.8	65.1	66.6	66.4	2.3	2.0	1.2	1.1
Legal highs	73.4	73.3	73.3	73.0	72.9	72.9	0.4	0.1	0.5	0.6
LSD	72.4	72.4	72.4	71.6	72.2	72.0	0.7	0.4	0.6	0.9
Methadone	69.3	69.5	69.4	68.6	69.2	69.1	1.2	1.4	1.0	1.1
M. mushrooms	71.0	70.9	71.0	70.0	70.6	70.4	1.2	0.9	1.1	0.9
Nicotine	66.3	66.0	66.2	66.1	65.2	65.8	0.9	0.7	1.0	1.1
VSA	70.4	70.5	70.5	67.8	70.4	70.1	3.0	1.7	1.9	0.8
Heroin pleiad	68.1	68.3	68.2	67.5	67.7	67.6	0.6	0.4	0.7	0.1
Ecstasy pleiad	76.2	75.5	76.0	75.8	75.2	75.6	1.1	0.3	0.2	0.5
Benz. pleiad	68.7	68.8	68.8	68.6	68.5	68.5	0.2	0.8	0.4	0.6
Illicit drugs	74.9	75.2	75.1	74.5	74.9	74.8	0.2	0.9	0.6	0.7

Table C.20. Performance and stability of linear discriminant for year-based definition of users (7 attributes).

	Total sample			LOOCV			Stability indicators			
Drug	Sn	Sp	Acc	Sn	Sp	Acc	TP→FN	FN→TP	FP→TN	TN→FP
Alcohol	58.0	58.1	58.0	57.7	54.4	57.5	4.7	5.3	3.7	5.9
Amphetamines	70.6	70.4	70.5	69.5	70.2	70.0	1.8	0.9	0.8	0.9
Amyl nitrite	65.4	65.2	65.3	63.2	65.0	64.8	2.3	3.0	2.7	3.7
Benz.	68.0	68.2	68.2	67.3	67.9	67.7	1.3	1.1	0.8	0.9
Cannabis	75.0	74.8	74.9	74.9	74.6	74.7	0.4	0.5	0.2	0.3
Chocolate	58.0	60.0	58.0	57.9	40.0	57.5	15.5	15.3	20.0	6.7
Cocaine	68.1	68.3	68.3	67.4	68.1	68.0	0.2	0.7	1.3	1.1
Caffeine	65.2	65.6	65.3	65.1	63.9	65.1	5.4	7.0	3.3	6.6
Crack	69.6	68.4	68.5	62.0	68.3	68.0	5.1	2.5	2.9	5.3
Ecstasy	70.6	70.1	70.2	69.8	70.0	69.9	1.0	0.4	0.4	0.8
Heroin	71.2	70.9	70.9	64.4	70.7	70.3	6.8	3.4	2.4	2.4
Ketamine	68.3	68.8	68.8	67.3	68.6	68.5	1.4	1.0	2.0	2.0
Legal highs	73.0	73.3	73.2	72.5	72.7	72.7	0.7	0.7	1.1	0.7
LSD	72.1	72.0	72.0	71.1	71.8	71.6	1.1	1.1	0.9	1.0
Methadone	69.1	69.3	69.3	68.1	69.1	68.9	0.9	2.2	1.0	1.4
M. mushrooms	71.7	71.8	71.8	71.0	71.7	71.5	0.9	0.9	0.8	0.9
Nicotine	67.5	67.8	67.6	67.3	67.3	67.3	0.3	0.5	0.8	0.4
VSA	68.4	68.5	68.5	63.2	68.3	68.1	5.3	5.3	2.8	3.5
Heroin pleiad	68.5	68.6	68.6	67.9	68.5	68.3	1.0	0.3	0.8	0.8
Ecstasy pleiad	75.9	76.1	76.0	75.7	75.6	75.6	0.4	0.6	0.9	0.4
Benz. pleiad	70.1	69.8	69.9	69.3	69.2	69.2	1.1	0.6	0.7	0.8
Illicit drugs	76.5	76.4	76.4	76.2	76.3	76.2	0.6	0.3	0.3	0.3

Table C.21. Performance and stability of linear discriminant for month-based definition of users (7 attributes).

	Total sample			LOOCV			Stability indicators			
Drug	Sn	Sp	Acc	Sn	Sp	Acc	TP→FN	FN→TP	FP→TN	TN→FP
Alcohol	58.5	57.8	58.4	57.8	56.0	57.5	2.4	1.7	3.3	2.4
Amphetamines	68.1	67.7	67.7	64.7	67.5	67.2	4.6	0.8	1.8	1.4
Amyl nitrite	70.7	70.8	70.8	61.0	70.6	70.4	2.4	2.4	3.9	6.7
Benz.	68.2	68.0	68.0	67.6	68.1	68.0	1.3	1.3	1.2	2.4
Cannabis	73.7	73.7	73.7	73.2	73.6	73.4	0.6	0.1	0.3	0.5
Chocolate	56.7	56.6	56.7	56.0	46.5	55.5	11.1	11.5	11.1	11.1
Cocaine	67.3	68.0	68.0	65.4	67.8	67.6	3.8	3.1	2.3	1.9
Caffeine	58.7	59.5	58.8	59.3	53.7	58.9	4.0	6.8	5.8	4.1
Crack	80.0	81.4	81.4	45.0	81.4	81.0	5.0	5.0	7.9	4.1
Ecstasy	66.7	66.6	66.6	65.4	66.5	66.4	1.7	1.7	1.6	1.6
Heroin	75.5	75.8	75.8	67.9	75.6	75.4	9.4	3.8	4.3	3.5
Ketamine	67.1	66.7	66.7	59.5	66.4	66.1	8.9	2.5	4.8	6.3
Legal highs	68.5	68.7	68.7	66.4	68.4	68.1	2.5	2.1	1.2	1.0
LSD	69.9	70.0	70.0	69.3	69.8	69.8	0.6	2.4	1.8	3.2
Methadone	66.7	66.9	66.9	63.7	66.6	66.4	2.9	2.3	2.1	1.6
M. mushrooms	71.7	71.4	71.4	67.9	71.1	70.9	4.4	3.1	1.3	1.4
Nicotine	64.1	64.2	64.1	63.5	63.7	63.6	1.0	0.7	0.8	0.5
VSA	70.6	70.3	70.3	55.9	70.2	69.9	11.8	5.9	4.2	7.8
Heroin pleiad	67.3	67.0	67.1	64.7	66.6	66.3	3.2	1.9	1.5	1.3
Ecstasy pleiad	74.4	74.4	74.4	74.0	74.2	74.1	0.3	0.3	0.4	0.8
Benz. pleiad	68.9	69.0	69.0	68.6	68.8	68.7	0.2	0.6	1.0	1.8
Illicit drugs	74.6	74.4	74.5	74.2	74.2	74.2	0.6	0.4	0.2	0.6

Table C.22. Performance and stability of linear discriminant for week-based definition of users (7 attributes).

	Total sample			LOOCV			Stability indicators			
Drug	Sn	Sp	Acc	Sn	Sp	Acc	TP→FN	FN→TP	FP→TN	TN→FP
Alcohol	56.0	55.6	55.9	55.2	54.1	54.9	1.4	1.3	1.6	2.4
Amphetamines	66.3	66.3	66.3	63.2	65.9	65.6	4.9	3.1	2.3	2.5
Amyl nitrite	70.6	85.0	84.8	47.1	84.9	84.5	0.0	5.9	13.1	4.6
Benz.	69.8	69.5	69.5	67.0	69.3	69.1	3.9	1.7	2.0	2.6
Cannabis	71.1	71.2	71.2	70.8	70.7	70.8	0.8	0.3	0.5	0.5
Chocolate	54.7	54.9	54.7	54.3	51.4	53.7	3.0	3.4	3.8	3.0
Cocaine	71.7	69.0	69.1	65.0	68.8	68.7	6.7	3.3	3.2	6.5
Caffeine	54.3	53.7	54.2	53.8	50.2	53.4	4.6	5.4	2.6	5.7
Crack	81.8	82.3	82.3	45.5	82.2	82.0	9.1	9.1	7.7	6.7
Ecstasy	70.2	71.4	71.4	65.5	71.2	70.9	7.1	6.0	3.8	3.3
Heroin	79.3	80.7	80.6	65.5	80.6	80.4	6.9	0.0	5.3	3.0
Ketamine	64.9	66.7	66.6	56.8	66.5	66.3	5.4	10.8	8.0	6.7
Legal highs	69.5	68.6	68.7	63.4	68.5	68.1	4.6	2.3	2.3	3.1
LSD	69.6	70.6	70.6	65.2	70.5	70.3	4.3	2.9	4.4	3.3
Methadone	68.6	68.0	68.1	64.5	67.9	67.7	5.8	2.5	2.7	4.0
M. mushrooms	72.7	72.9	72.9	63.6	72.8	72.6	6.8	2.3	5.5	4.9
Nicotine	62.5	62.7	62.6	62.1	62.3	62.2	0.4	1.2	0.9	0.9
VSA	71.4	74.3	74.3	61.9	74.2	74.1	9.5	4.8	8.3	5.5
Heroin pleiad	68.5	68.4	68.4	65.2	68.1	67.9	3.3	2.7	1.8	1.8
Ecstasy pleiad	72.3	72.5	72.4	72.0	72.1	72.0	0.5	0.1	0.5	0.4
Benz. pleiad	68.0	68.3	68.3	66.9	68.1	67.9	0.8	0.3	1.0	1.2
Illicit drugs	73.4	72.9	73.2	72.8	72.6	72.7	0.9	0.2	0.2	0.7

APPENDIX D

Minimal feature sets

D.1 Minimal feature sets for data set I (drug consumption) for heroin, ecstasy, and cannabis consumption.

In this appendix we show minimal feature sets for heroin, ecstasy, and cannabis consumption by LR and several classifiers. Symbol 'X' means used input feature. Results are calculated by LOOCV. column '#' contains number of used input features.

Table D.1. Minimal feature sets for heroin consumption based on ES selected by LR

Age	Edu	N	E	O	A	C	Imp	SS	Gndr	#	Model	Sn (%)	Sp (%)	Sum (%)
			X					X		2	M_1	69.81	67.36	137.18
		X					X		X	3	M_2	69.81	67.07	136.88
	X	X					X			3	M_3	66.98	66.89	133.87
X	X	X			X					4	M_4	69.81	66.71	136.52
	X				X		X		X	4	M_5	67.45	66.71	134.16
X				X		X			X	4	M_6	69.81	66.53	136.34
			X	X			X			3	M_7	66.51	66.59	133.10
	X			X			X			3	M_8	69.81	66.35	136.16
				X	X	X				3	M_9	67.93	66.35	134.27
					X	X	X		X	4	M_{10}	66.98	66.17	133.15
		X						X		2	M_{11}	70.28	66.11	136.39
			X		X	X	X			4	M_{12}	66.04	66.95	132.98
X		X					X			3	M_{13}	68.40	65.87	134.27
		X		X						2	M_{14}	66.51	65.81	132.32
X							X		X	3	M_{15}	68.40	65.75	134.15
				X	X	X			X	4	M_{16}	68.87	65.69	134.56
X	X		X	X						4	M_{17}	67.45	65.63	133.08
X			X				X			3	M_{18}	66.51	65.63	132.14
	X			X				X		3	M_{19}	66.04	65.57	131.61
		X				X	X			3	M_{20}	65.57	65.87	131.44
X						X	X			3	M_{21}	67.93	65.51	133.44
			X			X	X		X	4	M_{22}	65.57	65.45	131.02
X						X		X		3	M_{23}	66.51	65.39	131.90
				X		X		X		3	M_{24}	69.81	65.33	135.14
					X			X		2	M_{25}	68.87	65.33	134.20
X					X	X				3	M_{26}	67.45	65.33	132.79
X	X						X			3	M_{27}	66.98	65.33	132.31
X				X	X					3	M_{28}	67.45	65.27	132.73
		X			X	X			X	4	M_{29}	66.98	65.27	132.25
	X					X		X		3	M_{30}	69.81	65.21	135.02
			X		X		X		X	4	M_{31}	66.04	65.21	131.25
X					X		X			3	M_{32}	65.57	65.21	130.78
X		X							X	3	M_{33}	70.76	65.15	135.91
			X	X	X	X			X	5	M_{34}	69.34	65.15	134.49
								X	X	2	M_{35}	67.93	65.15	133.08
	X	X							X	3	M_{36}	67.93	65.15	133.08
X	X	X	X			X				5	M_{37}	66.04	65.15	131.19
	X				X	X	X			4	M_{38}	65.09	66.65	131.74
	X			X	X					3	M_{39}	65.09	66.53	131.62
				X		X	X			3	M_{40}	65.09	66.23	131.32
	X		X				X			3	M_{41}	65.09	65.93	131.02
	X			X		X				3	M_{42}	65.09	65.63	130.73
		X	X				X			3	M_{43}	65.09	65.51	130.61
X	X				X				X	4	M_{44}	69.34	65.09	134.43
X	X			X					X	4	M_{45}	68.40	65.09	133.49
	X					X	X		X	4	M_{46}	68.40	65.03	133.43
X			X	X		X				4	M_{47}	66.98	65.03	132.01

Table D.2. Minimal feature sets for cannabis consumption based on ES selected by LR

Age	Edu	N	E	O	A	C	Imp	SS	Gndr	#	Model	Sn (%)	Sp (%)	Sum (%)
X										1	M_1	73.12	67.90	141.03
	X						X		X	3	M_2	68.30	66.77	135.08
								X		1	M_3	66.32	75.32	141.65
	X					X	X			3	M_4	65.85	68.55	134.40
					X		X		X	3	M_5	66.40	65.48	131.89
		X					X		X	3	M_6	67.35	65.32	132.67
			X				X		X	3	M_7	65.61	65.32	130.94
				X						1	M_8	65.06	67.58	132.64
						X			X	2	M_9	65.69	65.00	130.69

Table D.3. Minimal feature sets for ecstasy consumption based on ES selected by LR

Age	Edu	N	E	O	A	C	Imp	SS	Gndr	Power	Symbol	Sn (%)	Sp (%)	Sum (%)
X								X		2	M_1	74.70	70.64	145.34
X				X						2	M_2	71.24	69.14	140.37
	X							X		2	M_3	68.98	68.78	137.76
X									X	2	M_4	68.71	72.31	141.02
						X		X		2	M_5	69.64	68.52	138.16
		X						X		2	M_6	68.98	68.34	137.32
					X			X		2	M_7	68.71	68.17	136.87
X					X					2	M_8	69.64	68.08	137.72
			X					X		2	M_9	68.04	68.52	136.56
				X				X		2	M_{10}	70.44	67.90	138.34
								X	X	2	M_{11}	70.04	67.20	137.24
X						X				2	M_{12}	69.51	67.20	136.70
X						X				2	M_{13}	67.11	71.78	138.89
X	X									2	M_{14}	72.70	67.02	139.72
				X		X				2	M_{15}	67.78	66.93	134.71
										3	M_{16}	67.11	66.40	133.51
	X	X	X			X	X		X	6	M_{17}	66.18	65.61	131.79
			X	X			X			3	M_{18}	65.51	65.52	131.03
				X			X		X	3	M_{19}	68.18	65.34	133.52
	X			X						2	M_{20}	67.51	65.26	132.77
		X		X			X			3	M_{21}	66.05	65.26	131.30
		X		X					X	3	M_{22}	67.78	65.17	132.94
X		X								2	M_{23}	69.11	65.08	134.19

Continued on the next page

Table D.5. *Continued*

Age	Edu	N	E	O	A	C	Imp	SS	Gndr	#	Classifier	Model	Sn (%)	Sp (%)	Sum (%)

Table D.5. Minimal feature sets for ecstasy consumption based on ES selected by several classifiers

Age	Edu	N	E	O	A	C	Imp	SS	Gndr	#	Classifier	Model	Sn (%)	Sp (%)	Sum (%)
X								X		2	PDFE	M1	72.44	72.40	144.84
X								X		2	GM	M1	72.70	72.05	144.75
X								X		2	KNN	M1	71.64	71.87	143.51
X								X		2	LR	M1	74.70	70.64	145.34
X								X		2	LDA	M1	77.50	68.43	145.93
X								X		2	NB	M1	65.38	77.69	143.07
X				X						2	KNN	M2	70.97	70.81	141.78
X				X						2	PDFE	M2	70.97	70.55	141.52
X				X						2	GM	M2	69.91	70.72	140.63
X				X						2	LR	M2	71.24	69.14	140.37
X				X						2	LDA	M2	66.05	72.31	138.36
	X							X		2	PDFE	M3	70.57	71.78	142.35
	X							X		2	PDFE	M3	70.57	71.78	142.35
	X							X		2	KNN	M3	71.37	70.02	141.39
	X							X		2	LR	M3	68.98	68.78	137.76
	X							X		2	GM	M3	68.04	69.40	137.44
						X		X		2	KNN	M4	70.31	70.55	140.85
						X		X		2	PDFE	M4	70.84	70.19	141.03
						X		X		2	GM	M4	69.11	69.40	138.51
						X		X		2	LR	M4	69.64	68.52	138.16
				X				X		2	PDFE	M5	70.17	70.55	140.72
				X				X		2	KNN	M5	70.17	70.19	140.37
				X				X		2	GM	M5	69.37	68.78	138.16
				X				X		2	LR	M5	70.44	67.90	138.34
X					X					2	KNN	M6	69.37	69.31	138.69
X					X					2	PDFE	M6	69.37	69.22	138.60
X					X					2	GM	M6	69.77	68.08	137.85

Continued on the next page

Table D.5. *Continued*

Age	Edu	N	E	O	A	C	Imp	SS	Gndr	#	Classifier	Model	Sn (%)	Sp (%)	Sum (%)
X					X					2	LR	M6	69.64	68.08	137.72
X					X					2	LDA	M6	71.77	66.76	138.53
					X			X		2	KNN	M7	69.37	69.22	138.60
					X			X		2	PDFE	M7	69.37	68.78	138.16
					X			X		2	LR	M7	68.71	68.17	136.87
					X			X		2	GM	M7	68.44	67.99	136.43
					X			X		2	LDA	M7	71.11	67.02	138.13
			X					X		2	KNN	M8	69.37	69.05	138.42
			X					X		2	PDFE	M8	68.98	69.05	138.02
			X					X		2	GM	M8	68.58	68.25	136.83
			X					X		2	LR	M8	68.04	68.52	136.56
			X					X		2	LDA	M8	65.38	71.43	136.81
		X						X		2	KNN	M9	69.64	68.96	138.60
		X						X		2	PDFE	M9	68.84	68.43	137.27
		X						X		2	LR	M9	68.98	68.34	137.32
		X						X		2	GM	M9	68.18	68.17	136.34
X									X	2	LR	M10	68.71	72.31	141.02
X									X	2	LDA	M10	68.71	72.31	141.02
X									X	2	GM	M10	68.71	72.31	141.02
X									X	2	KNN	M10	68.71	72.31	141.02
X									X	2	PDFE	M10	68.71	72.31	141.02
X									X	2	LR	M10	68.71	72.31	141.02
X							X			2	PDFE	M11	69.64	68.61	138.25
X							X			2	KNN	M11	71.11	67.90	139.01
X							X			2	GM	M11	67.78	70.11	137.88
X							X			2	LR	M11	72.70	67.02	139.72
X							X			2	LDA	M11	72.70	67.02	139.72
X						X				2	PDFE	M12	68.84	68.52	137.36
X						X				2	KNN	M12	68.84	68.17	137.01
X						X				2	GM	M12	67.78	68.70	136.47
X						X				2	LR	M12	69.51	67.20	136.70

Continued on the next page

Table D.5. *Continued*

Age	Edu	N	E	O	A	C	Imp	SS	Gndr	#	Classifier	Model	Sn (%)	Sp (%)	Sum (%)
X	X									2	PDFE	M13	68.04	69.22	137.27
X	X									2	LR	M13	67.11	71.78	138.89
X	X									2	LDA	M13	67.11	71.78	138.89
X	X									2	GM	M13	71.77	65.87	137.64
X	X									2	KNN	M13	65.78	73.90	139.68
							X	X		2	PDFE	M14	68.04	68.17	136.21
							X	X		2	KNN	M14	67.64	67.55	135.19
							X	X		2	GM	M14	70.31	65.96	136.27
	X			X						2	KNN	M15	67.91	68.08	135.99
	X			X						2	PDFE	M15	67.51	67.37	134.88
	X			X						2	GM	M15	66.84	66.40	133.25
	X			X						2	LR	M15	67.51	65.26	132.77
								X	X	2	KNN	M16	70.31	67.55	137.86
								X	X	2	PDFE	M16	70.31	67.55	137.86
								X	X	2	LR	M16	70.04	67.20	137.24
								X	X	2	GM	M16	70.04	67.20	137.24
				X		X				2	PDFE	M17	67.51	67.64	135.15
				X		X				2	GM	M17	67.38	67.37	134.75
				X		X				2	KNN	M17	67.51	67.28	134.79
				X		X				2	LR	M17	67.78	66.93	134.71
X			X							2	GM	M18	68.18	67.28	135.46
X			X							2	KNN	M18	66.98	66.49	133.47
X			X							2	PDFE	M18	66.31	66.67	132.98
X		X								2	GM	M19	67.24	67.20	134.44
X		X								2	PDFE	M19	67.51	67.11	134.62
X		X								2	KNN	M19	65.91	66.14	132.05
X		X								2	LR	M19	69.11	65.08	134.19
				X					X	2	PDFE	M20	66.31	65.96	132.27
	X						X			2	PDFE	M21	65.65	65.79	131.43
				X	X					2	GM	M22	66.05	65.61	131.65
				X	X					2	KNN	M22	65.78	65.43	131.21

Continued on the next page

Table D.5. *Continued*

Age	Edu	N	E	O	A	C	Imp	SS	Gndr	#	Classifier	Model	Sn (%)	Sp (%)	Sum (%)
				X	X					2	PDFE	M22	65.11	65.43	130.55
	X					X				2	PDFE	M23	66.05	65.52	131.57
						X	X			2	PDFE	M24	65.51	65.70	131.21
						X	X			2	KNN	M24	65.11	65.26	130.37
			X	X			X			3	LR	M25	65.51	65.52	131.03
			X	X			X			3	PDFE	M25	65.51	65.08	130.59
	X		X						X	3	PDFE	M26	65.65	65.34	130.99
	X				X				X	3	KNN	M27	65.51	65.34	130.86
	X				X				X	3	PDFE	M27	65.38	65.26	130.64
		X		X			X			3	LR	M28	66.05	65.26	131.30
		X		X			X			3	GM	M28	65.11	65.08	130.19
	X	X							X	3	KNN	M29	65.11	65.08	130.19

Table D.4. Minimal sets for cannabis consumption based on ES selected by several classifiers

Age	Edu	N	E	O	A	C	Imp	SS	Gndr	#	classifier	Model	Sn (%)	Sp (%)	Sum (%)
	X						X			2	PDFE	M1	68.93	68.87	137.80
	X						X			2	KNN	M1	68.62	69.68	138.29
X										1	GM	M2	73.12	67.90	141.03
X										1	KNN	M2	73.12	67.90	141.03
X										1	LDA	M2	73.12	67.90	141.03
X										1	LR	M2	73.12	67.90	141.03
X										1	PDFE	M2	73.12	67.90	141.03
	X							X		2	KNN	M3	67.59	67.58	135.17
	X							X		2	PDFE	M3	67.04	67.26	134.29
	X	X								2	KNN	M4	67.04	67.58	134.62
	X	X								2	PDFE	M4	65.85	66.45	132.30
							X	X		2	PDFE	M5	67.51	66.77	134.28
							X	X		2	KNN	M5	66.17	66.29	132.46
							X	X		2	GM	M5	66.72	65.65	132.37
					X		X			2	KNN	M6	66.40	66.45	132.86
						X				1	GM	M7	66.32	75.32	141.65
						X				1	KNN	M7	66.32	75.32	141.65
						X				1	LDA	M7	66.32	75.32	141.65
						X				1	LR	M7	66.32	75.32	141.65
						X				1	PDFE	M7	66.32	75.32	141.65
							X		X	2	GM	M8	65.85	67.26	133.11
							X		X	2	KNN	M8	65.85	67.26	133.11
							X		X	2	LDA	M8	65.85	67.26	133.11
							X		X	2	PDFE	M8	65.85	67.26	133.11
	X								X	2	PDFE	M9	65.85	66.13	131.98
	X				X					2	PDFE	M10	66.96	65.81	132.76
	X				X					2	KNN	M10	65.06	65.16	130.22
								X	X	2	PDFE	M11	66.72	65.65	132.37
				X						1	PDFE	M12	65.46	66.94	132.39
				X						1	GM	M12	65.06	67.58	132.64
				X						1	KNN	M12	65.06	67.58	132.64
				X						1	LDA	M12	65.06	67.58	132.64
				X						1	LR	M12	65.06	67.58	132.64
		X							X	2	PDFE	M13	65.30	65.16	130.46
	X		X							2	KNN	M14	65.06	65.81	130.87

D.2 Minimal feature sets for data set II (president elections of USA)

We illustrate the description for the second data set. We show the whole minimal feature sets for the second data base II for USA president elections by using one and several classifiers.

President Elections of USA

The dataset of president elections of USA consists of 31 instances. This dataset describe elections of the USA president elections from 1860 through 1980.There are several elections, in each election there are two basic opponents: the aspirant of the party currently in power (P-party) and the aspirant of the opposition party (O-party). There are 12 questions, the answers to these questions are 'yes', 'no' or 'unknown' [169].

As we can see this dataset has binary classes (P-party and O-party), and 12 Boolean features with identification year of president elections. These 12 questions are about the political, economic, social conditions of the country, and candidates themselves [4]. They are collected as follows:

1. Has the P-party been in power for more than one term?
2. Did the P-party receive more than 50% of the popular vote in the last election?
3. Was there significant activity of a third party during the election year?
4. Was there serious competition in the P-party primaries?
5. Was the P-party candidate the president at the time of the election?
6. Was there a depression or recession in the election year?
7. Was there a growth in the gross national product of more than 2.1% in the year of the election?
8. Did the P-party president make any substantial political changes during his term?
9. Did significant social tension exist during the term of the P-party?
10. Was the P-party administration guilty of any serious mistakes or scandals?

Table D.6. P-Party Victories

Election		Answer to questions											
#	year	1	2	3	4	5	6	7	8	9	10	11	12
p-1	1864	n	n	n	n	y	n	n	y	y	n	n	n
p-2	1868	y	y	n	n	n	n	y	y	y	n	y	n
p-3	1872	y	y	n	n	y	n	y	n	n	n	y	n
p-4	1880	y	n	n	y	n	n	y	y	n	n	n	n
p-5	1888	n	n	n	n	y	n	n	n	n	n	n	n
p-6	1900	n	y	n	n	y	n	y	n	n	n	n	y
p-7	1904	y	y	n	n	y	n	n	n	n	n	y	n
p-8	1908	y	y	n	n	n	n	n	y	n	n	n	y
p-9	1916	n	n	n	n	y	n	n	y	n	n	n	n
p-10	1924	n	y	y	n	y	n	y	y	n	y	n	n
p-11	1928	y	y	n	n	n	n	y	n	n	n	n	n
p-12	1936	n	y	n	n	y	y	y	y	n	n	y	n
p-13	1940	y	y	n	n	y	y	y	y	n	n	y	n
p-14	1944	y	y	n	n	y	n	y	y	n	n	y	n
p-15	1948	y	y	y	n	y	n	n	y	n	n	n	n
p-16	1956	n	y	n	n	y	n	n	n	n	n	y	n
p-17	1964	n	n	n	n	y	n	y	n	n	n	n	n
p-18	1972	n	n	n	n	y	n	y	y	y	n	n	n

11. Was the P-party candidate a national hero?

12. Was the O-party candidate a national hero?

Table D.6 shows the answers of all questions corresponding to the victories of the P-party. Table D.7 shows the answers of all questions corresponding to the victories of the O-party as follows:

Table D.7. O-Party Victories

Election		Answer to questions											
#	year	1	2	3	4	5	6	7	8	9	10	11	12
o-1	1860	y	n	y	y	n	n	y	n	y	n	n	n
o-2	1876	y	y	n	y	n	y	n	n	n	y	n	n
o-3	1884	y	n	n	y	n	n	y	n	y	n	y	n
o-4	1892	n	n	y	n	y	n	n	y	y	n	n	y
o-5	1896	n	n	n	y	n	y	n	y	y	n	y	n
o-6	1912	y	y	y	y	y	n	y	n	n	n	n	n
o-7	1920	y	n	n	y	n	n	n	y	y	n	n	n
o-8	1932	y	y	n	n	y	y	n	n	y	n	n	y
o-9	1952	y	n	n	y	n	n	y	n	n	y	n	y
o-10	1960	y	y	n	n	n	y	n	n	n	n	n	y
o-11	1968	y	y	y	y	n	n	y	y	y	n	n	n
o-12	1976	y	y	n	y	y	n	n	n	n	y	n	n
o-13	1980	n	n	y	y	y	y	n	n	n	y	n	y

Table D.8. Minimal sets for president elections of USA based on ES selected by LR

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Symbol	power	Sn %	Sp %	Sum %
		X		X			X	X				M_1	4	92	89	181
X	X	X			X							M_2	4	92	83	176
X		X			X			X				M_3	4	85	83	168
X		X					X	X				M_4	4	92	78	170
X							X	X		X		M_5	4	85	78	162
X								X	X			M_6	3	85	78	162
	X	X		X		X	X					M_7	5	85	78	162
		X			X		X	X				M_8	4	77	100	177
			X									M_9	1	77	94	171
X	X								X		X	M_{10}	4	77	83	160
X					X			X			X	M_{11}	4	77	83	160
X	X	X									X	M_{12}	4	77	83	160
X				X	X	X	X		X			M_{13}	6	77	83	160
X	X				X				X			M_{14}	4	77	83	160
		X		X	X		X					M_{15}	4	77	83	160
X				X		X	X		X		X	M_{16}	6	77	78	155
X		X							X		X	M_{17}	4	77	78	155
X							X	X			X	M_{18}	4	77	78	155
		X				X	X	X				M_{19}	4	77	78	155
X	X					X	X	X				M_{20}	5	77	78	155
	X	X			X		X					M_{21}	4	77	78	155
X		X				X						M_{22}	3	77	78	155

Continued on the next page

Table D.9. *Continued*

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	classifiers	Model	#	Sn (%)	Sp (%)	Sum (%)

Table D.9. Minimal feature sets for president elections based on ES by several classifiers

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	classifiers	Model	#	Sn (%)	Sp (%)	Sum (%)
X					X			X			X	NB	M1	4	85	83	168
X						X				X	X	KNN	M2	4	85	83	168
X					X			X			X	KNN	M3	4	85	83	168
	X	X				X				X		KNN	M4	4	85	83	168
X	X	X							X			KNN	M5	4	85	83	168
								X	X			GM	M6	2	85	83	168
		X								X	X	KNN	M7	3	100	78	178
		X							X		X	KNN	M8	3	100	78	178
		X			X							KNN	M9	2	100	78	178
								X	X			LR	M10	2	85	78	162
								X	X			NB	M11	2	85	78	162
								X	X			LDA	M12	2	85	78	162
				X		X					X	KNN	M13	3	85	78	162
X	X									X		KNN	M14	3	85	78	162
								X	X			KNN	M15	2	85	78	162
	X			X	X			X				KNN	M16	4	85	78	162
			X									GM	M17	1	77	100	177
			X									LR	M18	1	77	94	171
			X									NB	M19	1	77	94	171
			X									LDA	M20	1	77	94	171
X		X					X					KNN	M21	3	77	94	171
			X									KNN	M22	1	77	94	171
X								X		X	X	KNN	M23	4	77	89	166
X	X								X		X	KNN	M24	4	77	89	166
	X	X							X	X		KNN	M25	4	77	89	166
X	X				X				X			KNN	M26	4	77	89	166

Continued on the next page

Table D.9. *Continued*

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	classifiers	Model	#	Sn (%)	Sp (%)	Sum (%)
X	X								X		X	LR	M27	4	77	83	160
X					X			X			X	LR	M28	4	77	83	160
X	X	X									X	LR	M29	4	77	83	160
X	X				X				X			LR	M30	4	77	83	160
X		X						X			X	NB	M31	4	77	83	160
X	X	X									X	LDA	M32	4	77	83	160
				X				X		X	X	KNN	M33	4	77	83	160
	X				X					X	X	KNN	M34	4	77	83	160
	X				X		X		X		X	KNN	M35	5	77	83	160
X						X		X			X	KNN	M36	4	77	83	160
X		X						X			X	KNN	M37	4	77	83	160
	X							X			X	KNN	M38	3	77	83	160
X	X				X						X	KNN	M39	4	77	83	160
X	X	X									X	KNN	M40	4	77	83	160
		X				X	X		X	X		KNN	M41	5	77	83	160
					X			X		X		KNN	M42	3	77	83	160
					X		X	X				KNN	M43	3	77	83	160
					X	X		X				KNN	M44	3	77	83	160
X					X	X	X					KNN	M45	4	77	83	160
	X			X		X	X					KNN	M46	4	77	83	160
				X	X	X						KNN	M47	3	77	83	160
		X		X		X						KNN	M48	3	77	83	160
X		X				X						KNN	M49	3	77	83	160
X	X								X		X	GM	M50	4	77	83	160
X	X				X				X			GM	M51	4	77	83	160
		X							X		X	LR	M52	3	77	78	155
							X	X			X	LR	M53	3	77	78	155
X		X				X						LR	M54	3	77	78	155
		X							X		X	NB	M55	3	77	78	155
X		X				X						NB	M56	3	77	78	155
				X	X	X						LDA	M57	3	77	78	155

Continued on the next page

Table D.9. *Continued*

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	classifiers	Model	#	Sn (%)	Sp (%)	Sum (%)
X		X				X						LDA	M58	3	77	78	155
					X	X	X		X	X	X	KNN	M59	6	77	78	155
	X					X			X	X	X	KNN	M60	5	77	78	155
	X							X		X	X	KNN	M61	4	77	78	155
				X			X			X	X	KNN	M62	4	77	78	155
							X	X			X	KNN	M63	3	77	78	155
X				X				X			X	KNN	M64	4	77	78	155
X	X					X					X	KNN	M65	4	77	78	155
				X	X						X	KNN	M66	3	77	78	155
		X		X							X	KNN	M67	3	77	78	155
	X			X							X	KNN	M68	3	77	78	155
				X		X	X			X		KNN	M69	4	77	78	155
	X						X			X		KNN	M70	3	77	78	155
				X	X					X		KNN	M71	3	77	78	155
X					X					X		KNN	M72	3	77	78	155
X				X						X		KNN	M73	3	77	78	155
						X	X	X				KNN	M74	3	77	78	155
				X			X	X				KNN	M75	3	77	78	155
	X	X					X	X				KNN	M76	4	77	78	155
X							X	X				KNN	M77	3	77	78	155
	X			X		X		X				KNN	M78	4	77	78	155
X	X					X	X					KNN	M79	4	77	78	155
				X	X		X					KNN	M80	3	77	78	155
X	X				X		X					KNN	M81	4	77	78	155
		X		X			X					KNN	M82	3	77	78	155
	X	X		X								KNN	M83	3	77	78	155
X									X	X		GM	M84	3	77	78	155
				X					X			GM	M85	2	77	78	155
X		X				X						GM	M86	3	77	78	155

Bibliography

[1] Fehrman E, ..., Muhammad AK. Personality Traits and Drug Consumption: A Story Told by Data. Springer, 2018, to be published (accepted).

[2] Fehrman E, Muhammad AK, Mirkes EM, Egan V, Gorban AN. The Five Factor Model of personality and evaluation of drug consumption risk. In Francesco Palumbo et al. (eds.), Data Science, Studies in Classification, Data Analysis, and Knowledge Organization, Springer, 2017, pp. 215-226. (Extended technical report in ArXiv preprint, 2015. https://arxiv.org/abs/1506.06297)

[3] Fehrman E, Egan V. Drug consumption, collected online March 2011 to March 2012, English-speaking countries. ICPSR36536-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-09-09. Deposited by Mirkes E. http://doi.org/10.3886/ICPSR36536.v1

[4] Lichtman AJ, Keilis-Borok VI. Pattern Recognition as Applied to Presidential Elections in USA, 1860-1980. Role of Integral Social, Economic and Political Traits, Contribution Division of Geological and Planetary Sciences, California Institute of Technology, 1981(3760).

[5] Wolberg WH. Breast cytology diagnosis via digital image analysis. Analytical and Quantitative Cytology and Histology; 1993; 15(6): 396-404.

[6] Kleiman MA, Caulkins JP, Hawken A. Drugs and Drug Policy: What Everyone Needs to Know. Oxford University Press; 2011.

[7] Beaglehole R, Bonita R, Horton R, Adams C, Alleyne G, Asaria P, Baugh V, Bekedam H, Billo N, Casswell S, et al. Priority actions for the non-communicable disease crisis. The Lancet. 2011; 377(9775):1438–1447.

[8] Bickel WK, Johnson MW, Koffarnus MN, MacKillop J, Murphy G James. The behavioral economics of substance use disorders: reinforcement pathologies and their repair. Annual review of clinical psychology. 2014; 10:641–677.

[9] McGinnis JM, Foege WH. Actual causes of death in the United States. Journal of the American Medical Association. 1993; 270(18):2207–2212.

[10] Sutina AR, Evans MK, Zonderman AB. Personality traits and illicit substances: the moderation role of poverty. Drug and Alcohol Dependence. 2013; 131:247–251.

[11] Cleveland MJ, Feinberg ME, Bontempo DE, Greenberg MT. The role of risk and protective factors in substance use across adolescence. Journal of Adolescent Health. 2008; 43(2):157–164.

[12] Ventura CA, de Souza J, Hayashida M, Ferreira PS. Risk factors for involvement with illegal drugs: opinion of family members or significant others. Journal of Substance Use. 2014; 20(2):136–142.

[13] WHO. Prevention of mental disorders: Effective interventions and policy options: Summary report. Geneva: World Health Organization; 2004. Available: http://www.who.int/mental_health/evidence/en/prevention_of_mental_disorders_sr.pdf.

[14] Dubey C, Arora M, Gupta S, Kumar B. Five factor correlates: A comparison of substance abusers and non-substance abusers. Journal of the Indian Academy of Applied Psychology. 2010; 36(1):107–114.

[15] Bogg T, Roberts BW. Conscientiousness and health-related behaviors: A meta-analysis of the leading behavioral contributors to mortality. Psychological Bulletin. 2004; 130(6):887.

[16] Khantzian EJ, Khantzian NJ. Cocaine Addiction: Is There a Psychological Predisposition?. Psychiatric Annals. 1984 Oct 1;14(10):753-9.

[17] Goldberg LR. The structure of phenotypic personality traits. American psychologist. 1993 Jan;48(1):26–34.

[18] Thurstone LL. The vectors of mind. Psychological review. 1934 Jan;41(1):1-32.

[19] Digman JM. Personality structure: Emergence of the five-factor model. Annual review of psychology. 1990 Feb;41(1):417–40.

[20] Costa PT, MacCrae RR. Revised NEO-Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional manual. Odessa, FL: Psychological Assessment Resources; 1992.

[21] McCrae RR, John OP. An introduction to the fiveâĂŘfactor model and its applications. Journal of personality. 1992 Jun 1;60(2):175-215.

[22] Roncero C, Daigre C, Barral C, Ros-Cucurull E, Grau-López L, Rodríguez-Cintas L, Tarifa N, Casas M, Valero S. Neuroticism associated with cocaine-induced psychosis in cocaine-dependent patients: a cross-sectional observational study. PloS one. 2014; 9(9):e106,111.

[23] Vollrath M, Torgersen S. Who takes health risks? a probe into eight personality types. Personality and Individual Differences. 2002; 32(7):1185–1197.

[24] Flory K, Lynam D, Milich R, Leukefeld C, Clayton R. The relations among personality, symptoms of alcohol and marijuana abuse, and symptoms of comorbid psychopathology: results from a community sample. Experimental and Clinical Psychopharmacology. 2002; 10(4):425–434.

[25] Andreassen CS, Griffiths MD, Gjertsen SR, Krossbakken E, Kvam S, Pallesen S. The relationships between behavioral addictions and the five-factor model of personality. Journal of Behavioral Addictions. 2013; 2(2): 90–99.

[26] Terracciano A, Lóckenhoff CE, Crum RM, Bienvenu OJ, Costa PT. Five-Factor Model personality profiles of drug users. Bmc Psychiatry. 2008; 8(1):22.

[27] Turiano NA, Whiteman SD, Hampson SE, Roberts BW, Mroczek DK. Personality and substance use in midlife: Conscientiousness as a moderator and the effects of trait change. Journal of research in personality. 2012; 46(3):295–305.

[28] Stewart SH, Devine H. Relations between personality and drinking motives in young adults. Personality and Individual Differences. 2000; 29(3):495–511.

[29] Haider AH, Edwin DH, MacKenzie EJ, Bosse MJ, Castillo RC, Travison TG, Group LS, et al. The use of the NEO-Five Factor inventory to assess personality in trauma patients: a two-year prospective study. Journal of Orthopaedic trauma. 2002; 16(9):660–667.

[30] Belcher AM, Volkow ND, Moeller FG, Ferré S, Personality traits and vulnerability or resilience to substance use disorders, Trends in cognitive sciences. 2016; 18(4):2

[31] Kopstein AN, Crum RM, Celentano DD, Martin SS. Sensation seeking needs among 8th and 11th graders: characteristics associated with cigarette and marijuana use. Drug and alcohol dependence. 2001; 62(3):195–203.

[32] Yasnitskiy L, Gratsilev V, Kulyashova J, Cherepanov F. Possibilities of artificial intellect in detection of predisposition to drug addiction. Perm University Herald Series “Philosophy Psychology Sociology”. 2015; 1(21):61–73.

[33] Valero S, Daigre C, Rodríguez-Cintas L, Barral C, Gomà-i-Freixanet M, Ferrer M, Casas M, Roncero C. Neuroticism and impulsivity: Their hierarchical organization in the personality characterization of drug-dependent patients from a decision tree learning perspective. Comprehensive Psychiatry. 2014; 55(5):1227–1233.

[34] Bulut F, Bucak IÖ. An urgent precaution system to detect students at risk of substance abuse through classification algorithms. Turkish Journal of Electrical Engineering & Computer Sciences. 2014; 22(3):690–707.

[35] Rioux C, Castellanos-Ryan N, Parent S, Ségu JR, The interaction between temperament and the family environment in adolescent substance use and externalizing behaviors: Support for diathesisâĂŞstress or differential susceptibility? Developmental Review. 2016; 40: 117–150.

[36] Weissman DG, Schriber RA, Fassbender C, Atherton O, Krafft C, Robins RW, Hastings PD, Guyerb AE, Earlier adolescent substance use onset predicts stronger connectivity between reward and cognitive control brain networks. Developmental Cognitive Neuroscience. 2015; 16:121–129.

[37] McCrae RR, Costa PT. A contemplated revision of the NEO Five-Factor Inventory. Personality and Individual Differences. 2004; 36(3):587–596.

[38] Stanford MS, Mathias CW, Dougherty DM, Lake SL, Anderson NE, Patton JH. Fifty years of the Barratt Impulsiveness Scale: An update and review. Personality and Individual Differences. 2009; 47(5):385–395.

[39] Zuckerman M. Behavioral expressions and biosocial bases of sensation seeking. New York: Cambridge University Press; 1994.

[40] Bruinsma K, Taren DL. Chocolate: food or drug? Journal of the American Dietetic Association. 1999; 99(10):1249–1256.

[41] 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. Personality and Individual Differences. 2000; 29(5):907–920.

[42] Gurrera RJ, Nestor PG, O'Donnell BF. Personality traits in schizophrenia: comparison with a community sample. The Journal of Nervous and Mental Disease. 2000; 188(1):31–35.

[43] Terentjev PV. Biometrische Untersuchungen ′Uber Die Morpho-Logischen Merkmale Von Rana Ridibunda Pall:(Amphibia, Salientia). Biometrika. 1931; 1:23–51. Available: http://www.jstor.org/stable/2333629

[44] Mitteroecker P, Bookstein F. The conceptual and statistical relationship between modularity and morphological integration. Systematic biology. 2007; 56(5):818-836.

[45] Berg RL. The ecological significance of correlation pleiades. Evolution. 1960; 1:171-180. Available: http://www.jstor.org/stable/2405824

[46] Armbruster WS, Di Stilio VS, Tuxill JD, Flores TC, Runk JL. Covariance and decoupling of floral and vegetative traits in nine Neotropical plants: a re-evaluation of Berg's correlation-pleiades concept. American Journal of Botany. 1999; 86(1):39–55.

[47] Krishnapuram R, Keller JM. A possibilistic approach to clustering. IEEE transactions on fuzzy systems. 1993; 1(2):98–110.

[48] Bezdek, J. Pattern recognition with fuzzy objective function algorithms. New York, NY: Plenum Press; 1981.

[49] Xu R, Wunsch D. Clustering. Hoboken, New Jersey: John Wiley & Sons Inc.; 2008.

[50] Omote H, Sugiyama K. Method for drawing intersecting clustered graphs and its application to web ontology language. In Proceed-

ings of the 2006 Asia-Pacific Symposium on Information Visualisation-Volume 60 2006 Jan 1 (pp. 89-92). Australian Computer Society, Inc. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.294.2209&rep=rep1&type=pdf

[51] Tang JA. Feature selection for classification: A review. In S. A. Jiliang Tang, Data Classification: Algorithms and Applications; 2014; 37.

[52] Yang H, Moody J. Feature selection based on joint mutual information. in: Proceedings of International ICSC Symposium on Advances in Intelligent Data Analysis; 1999: 22–25.

[53] Luo LY. Methods of forward feature selection based on the aggregation of classifiers generated by single attribute. Computers in biology and medicine; 2011; 41(7): 435–441.

[54] Guyon I. An introduction to variable and feature selection. The Journal of Machine Learning Research; 2003; 3: 1157-1182.

[55] Kohavi R. Wrappers for feature subset selection. Artificial intelligence; 1997; 97(1): 273-324.

[56] Dash M. Feature selection for classification. Intelligent data analysis; 1997; 1(1): 131-156.

[57] John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. John GH, Kohavi R and Pfleger K, 1994. Irrelevant fea In Machine learning: proceedings of the eleventh international conference; 1994: 121-129.

[58] MacWilliams FJ, Sloane NJ. The theory of error correcting codes. Elsevier 1977.

[59] Dietterich TG. Ensemble methods in machine learning. InInternational workshop on multiple classifier systems. Springer Berlin Heidelberg; 2000: 1-15.

[60] Sugumaran V. Insights into advancements in intelligent information technologies : discoveries. Hershey, PA: Information Science Reference; 2012.

[61] Wolberg WH, Street WN, Mangasarian OL. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer letters; 1994;77(2):163-71.

[62] Hutchinson ML, Isenstein LM, Zahniser DJ. high-resolusion and contextual analysis for the diagnosis of fine needle aspirations of breast. Anal Quant Cytol Histol;1991; 13: 351-355.

[63] Surveygizmo: Professional Survey Solution. https://www.surveygizmo.com/ accessed 30/04/2017.

[64] Bhaskaran V, LeClaire J. Online surveys for dummies. John Wiley & Sons; 2010.

[65] 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

[66] Atkinson R, Flint J. Accessing hidden and hard-to-reach populations: Snowball research strategies. Social research update. 2001 Jan;33(1):1-4.

[67] Salganik MJ, Heckathorn DD. Sampling and estimation in hidden populations using respondentâĂŘdriven sampling. Sociological methodology. 2004 Dec 1;34(1):193-240.

[68] Kaplan CD, Korf D, Sterk C. Temporal and Social Contexts of Heroin-Using Populations An Illustration of the Snowball Sampling Technique. The Journal of nervous and mental disease. 1987 Sep 1;175(9):566-74.

[69] Lopes CS, Rodrigues LC, Sichieri R. The lack of selection bias in a snowball sampled case-control study on drug abuse. International journal of epidemiology. 1996 Dec 1;25(6):1267-70.

[70] McKnight C, Des Jarlais D, Bramson H, Tower L, Abdul-Quader AS, Nemeth C, Heckathorn D. Respondent-driven sampling in a study of drug users in New York City: notes from the field. Journal of Urban Health. 2006 Nov 1;83(1):54.

[71] Egan V. Individual differences and antisocial behaviour. In: Furnham A, Stumm S, Petredies K, editors. Handbook of Individual Differences. Oxford: Blackwell-Wiley; 2011. pp. 512–537.

[72] McCrae RR, Costa PT. The NEO Personality Inventory: Using the Five-Factor model in counseling. Journal of Counseling & Development. 1991; 69(4):367–372.

[73] Jakobwitz S, Egan V. The dark triad and normal personality traits. Personality and Individual Differences. 2006; 40(2):331–339.

[74] McHoskey JW, Worzel W, Szyarto C. Machiavellianism and psychopathy. Journal of personality and social psychology. 1998 Jan;74(1):192–210.

[75] Raskin RN, Hall CS. A narcissistic personality inventory. Psychological reports. 45(2), Oct 1979, 590.

[76] Settles RE, Fischer S, Cyders MA, Combs JL, Gunn RL, Smith GT. Negative urgency: a personality predictor of externalizing behavior characterized by neuroticism, low conscientiousness, and disagreeableness. Journal of Abnormal Psychology. 2012; 121(1):160–172.

[77] Snowden RJ, Gray NS. Impulsivity and psychopathy: Associations between the barrett impulsivity scale and the psychopathy checklist revised. Psychiatry Research. 2011; 187(3):414–417.

[78] García-Montes JM, Zaldívar-Basurto F, López-Ríos F, Molina-Moreno A. The role of personality variables in drug abuse in a Spanish university population. International journal of mental health and addiction. 2009; 7(3):475–487.

[79] Fossati P, Ergis AM, Allilaire JF. Problem-solving abilities in unipolar depressed patients: comparison of performance on the modified version of the Wisconsin and the California sorting tests. Psychiatry Research. 2001; 104(2):145–156.

[80] McMurran M, Blair M, Egan V. An investigation of the correlations between aggression, impulsiveness, social problemâĂŘsolving, and alcohol use. Aggressive Behavior. 2002 Jan 1;28(6):439-45.

[81] McDaniel SR, Mahan JE. An examination of the Impss scale as a valid and reliable alternative to the SSS-V in optimum stimulation level research. Personality and Individual Differences. 2008; 44(7):1528–1538.

[82] Ragan DT, Beaver KM. Chronic Offenders: A Life-Course Analysis of Marijuana Users. Youth & Society. 2010; 42:174–198.

[83] HomeOfficeUK. Drug misuse:findings from the 2013 to 2014 CSEW 2nd ed. 2014. Available: https://www.gov.uk/government/statistics/drug-misuse-findings-from-the-2013-to-2014-csew.

[84] Wright KB. Researching InternetâĂŘbased populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. Journal of ComputerâĂŘMediated Communication. 2005 Apr 1;10(3): http://doi.org/10.1111/j.1083-6101.2005.tb00259.x.

[85] Fridberg DJ, Vollmer JM, O'Donnell BF, Skosnik PD. Cannabis users differ from non-users on measures of personality and schizotypy. Psychiatry Research. 2011; 186(1):46–52.

[86] Lee SY, Poon WY, Bentler PM. A two-stage estimation of structural equation models with continuous and polytomous variables. British Journal of Mathematical and Statistical Psychology. 1995; 48(2):339–358.

[87] Martinson EO, Hamdan MA. Maximum likelihood and some other asymptotically efficient estimators of correlation in two way contingency tables. Journal of Statistical Computation and Simulation. 1971; 1(1):45–54.

[88] Linting M, van der Kooij A. Nonlinear Principal Components Analysis with CATPCA: A tutorial. Journal of Personality Assessment. 2012; 94(1):12–25.

[89] Gorban AN, Zinovyev AY. Principal graphs and manifolds. In: Olivas ES, Guerrero JDM, Sober MM, Benedito JRM, López AJS, editors. Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. Hershey – New York. IGI Global; 2009. pp. 28–59.

[90] Gorban AN, Zinovyev AY. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. International journal of neural systems. 2010; 20(03):219–232.

[91] Gorban AN, Kégl B, Wunsch DC, Zinovyev AY (eds). Principal Manifolds for Data Visualisation and Dimension Reduction. LNCSE, Vol. 58, Berlin-Heidelberg-New York. Springer; 2008.

[92] Pearson K. On lines and planes of closest fit to system of points in space. Philosophical magazine. 1901; 2(6):559–572.

[93] Guttman L. Some necessary conditions for common-factor analysis. Psychometrika. 1954; 19(2):149–161.

[94] Kaiser HF. The application of electronic computers to factor analysis. Educational and Psychological Measurement. 1960; 20:141–151.

[95] Gujarati DN. Basic econometrics, 4th ed. New York: McGraw-Hill; 2003.

[96] McCabe GP. Principal variables. Technometrics. 1984; 26(2):137–144.

[97] Naikal N, Yang AY, Sastry SS. Informative feature selection for object recognition via Sparse PCA. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE; 2011. pp. 818–825. doi: 10.1109/ICCV.2011.6126321

[98] Clarkson KL. Nearest-neighbor searching and metric space dimensions. In: Shakhnarovich G, Darrell T, Indyk P, editors. Nearest-neighbor methods for Learning and Vision: Theory and Practice. MIT Press; 2006. pp. 15–59.

[99] Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936; 7(2):179–188.

[100] Hastie T, Tibshirani R. Discriminant adaptive nearest neighbor classification. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1996; 18(6):607–616.

[101] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y., Zhou, Z.H.: Top 10 algorithms in data mining. Knowl. Inf. Syst. 2008; 14(1), 1-37.

[102] Li Q, Racine JS. Nonparametric econometrics: theory and practice. Princeton University Press; 2007.

[103] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Belmont, Calif: Wadsworth International Group; 1984.

[104] Quinlan JR. Simplifying decision trees. International Journal of Man-Machine Studies. 1987; 27(3):221–234.

[105] Rokach L, Maimon O. Decision trees. In: Rokach L, Maimon O, editors. Data Mining and Knowledge Discovery Handbook. Berlin: Springer; 2010. pp. 165–192.

[106] Sofeikov KI, Tyukin IY, Gorban AN, Mirkes EM, Prokhorov DV, Romanenko IV. Learning optimization for decision tree classification of non-categorical data with information gain impurity criterion. In: Neural Networks (IJCNN); 2014. Interna. Joint Confe. on, IEEE; 2014. pp. 3548–3555.

[107] Gelfand SB, Ravishankar CS, Delp EJ. An iterative growing and pruning algorithm for classification tree design. IEEE Transaction on Pattern Analysis and Machine Intelligence. 1991; 13(2):163–174.

[108] Dietterich T, Kearns M, Mansour Y. Applying the weak learning framework to understand and improve C4.5. In: ICML, Proc. of the 13th Int. Conf. on Machine Learning. San Francisco: Morgan Kaufmann; 1996. pp. 96–104.

[109] Kearns M, Mansour Y. On the boosting ability of top-down decision tree learning algorithms. Journal of Computer and System Sciences. 1999; 58(1):109–128.

[110] Dinov ID. Expectation Maximization and Mixture Modeling Tutorial. UCLA, Statistics Online Computational Resource; 2008. Available: http://escholarship.org/uc/item/1rb70972.

[111] Buhmann MD. Radial basis functions: theory and implementations, vol 12. Cambridge: University Press; 2003.

[112] Scott DW. Multivariate Density Estimation: Theory, Practice, and Visualization, 1st edn. New York: Wiley; 1992.

[113] Hosmer Jr DW, Lemeshow S. Applied logistic regression. John Wiley & Sons; 2004.

[114] Russell S, Norvig P. Artificial intelligence: a modern approach; 1995.

[115] Breiman L. Random Forests. Machine Learning. 2001; 45(1):5–32.

[116] Biau G. Analysis of a random forests model. The Journal of Machine Learning Research. 2012; 13(1):1063–1095.

[117] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York: Springer; 2009.

[118] Liaw A, Wiener MR. Classification and Regression by randomForest. R news. 2002; 2(3):18–22.

[119] Williams G. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Springer Science & Business Media; 2011.

[120] Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics surveys. 2010; 4:40–79.

[121] Gorban A, Zinovyev A. Elastic principal graphs and manifolds and their practical applications. Computing. 2005 Aug 18;75(4):359–79.

[122] Hastie T, Stuetzle W. Principal curves. Journal of the American Statistical Association. 1989 Jun 1;84(406):502–16.

[123] Gorban AN, Zinovyev AY. Principal graphs and manifolds. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques. Emilio Soria Olivas et al. (eds), IGI Global, Hershey, PA, USA, 2009, pp. 28–59.

[124] Jolliffe, I.T.: Principal Component Analysis. Springer, New York. 1986. https://doi.org/10.1007/978-1-4757-1904-8

[125] Jolliffe, I.T., Trendafilov, N.T., Uddin, M.: A modified principal component technique based on the LASSO. J. comput. Graph. Stat. 12(3), 531–547 (2003).

[126] Kwak, N.: Principal component analysis based on L1-norm maximization. IEEE trans. pattern anal. mach. intell. 30(9), 1672–1680 2008.

[127] Brooks, J., Dulá, J., Boone, E.: A pure L1-norm principal component analysis, Comput. Stat. Data Anal. 61, 83–98, 2013.

[128] Gorban, A.N., Mirkes, E.M., Zinovyev, A.: Piece-wise quadratic approximations of arbitrary error functions for fast and robust machine learning. Neural Netw. 84, 28–38, 2016.

[129] Chacón M, Lévano M, Allende H, Nowak H. Detection of gene expressions in microarrays by applying iteratively elastic neural net. Adaptive and Natural Computing Algorithms. 2007:355–63.

[130] Zinovyev A. Data visualization in political and social sciences, In International Encyclopedia of Political Science, Badie, B., Berg-Schlosser, D., Mor-

lino, L. A. (Eds.), SAGE 2011. arXiv e-print http://arxiv.org/abs/1008.1188

[131] Resta M. Computational intelligence paradigms in economic and financial decision making. Springer; 2016.

[132] Shaban H, Tavoularis S. Identification of flow regime in vertical upward airâĂŞwater pipe flow using differential pressure signals and elastic maps. International Journal of Multiphase Flow. 2014 May 31;61:62–72.

[133] Becker P. Recognitions of patterns. Polyteknisk, Copenhagen. 1968.

[134] Fisher RA. The use of multiple measurements in taxonomic problems. Annals of eugenics. 1936 Sep 1;7(2):179-188.

[135] McCrae RR, Costa Jr PT, Terracciano A, Parker WD, Mills CJ, De Fruyt F, Mervielde I. Personality trait development from age 12 to age 18: longitudinal, cross-sectional, and cross-cultural analyses. Journal of Personality and Social Psychology, 2002; 83(6): 1456–1468.

[136] Shock NW, Greulich RC, Andres R, Arenberg D, Costa Jr PT. Lakatta, EG, Tobin, JD (1984). Normal human aging: The Baltimore Longitudinal Study of Aging (NIH Publication No. 84-2450). Bethesda, MD: National Institutes of Health.

[137] Grossman JC, Goldstein R, Eisenman R. Undergraduate marijuana and drug use as related to openness to experience. Psychiatric Quarterly. 1974 Mar 1;48(1):86–92.

[138] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. 1995; 57(1):289–300.

[139] Mitchell TM. Machine learning. 1997. Burr Ridge, IL: McGraw Hill 45; 1997.

[140] Ellis, PD. The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results. United Kingdom: Cambridge University Press, 2010.

[141] Blockeel H, Struyf J. Efficient algorithms for decision tree cross-validation. Journal of Machine Learning Research. 2002;3(Dec):621–650.

[142] Hoadley B. [Statistical Modeling: The Two Cultures]: Comment. Statistical Science. 2001 Aug 1;16(3):220–224.

[143] Hand DJ. Classifier technology and the illusion of progress. Statistical science. 2006;21(1):1-4.

[144] Belsley DA, Kuh E, Welsch RE. Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley & Sons; 2005.

[145] Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, CarrÃľ G, MarquÃľz JR, Gruber B, Lafourcade B, LeitÃčo PJ, MÃijnkemÃijller T. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 2013 Jan 1;36(1):27–46.

[146] Gorban AN, Zinovyev A. Fast and user-friendly non-linear principal manifold learning by method of elastic maps. In Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on 2015 Oct 19 (pp. 1-9). IEEE.

[147] Mirkes EM, Alexandrakis I, Slater K, Tuli R, Gorban AN. Computational diagnosis and risk evaluation for canine lymphoma. Computers in Biology and Medicine. 2014a; 53:279–290.

[148] Mirkes EM, Alexandrakis I, Slater K, Tuli R, Gorban AN. Computational diagnosis of canine lymphoma. Journal of Physics: Conference Series. 2014; 490(1):012135. Available: http://stacks.iop.org/1742-6596/490/i=1/a=012135.

[149] Marks S, Dunn OJ. Discriminant functions when covariance matrices are unequal. Journal of the American Statistical Association. 1974 Jun 1;69(346):555–559.

[150] Gorban AN, Zinovyev AY. Elastic maps and nets for approximating principal manifolds and their application to microarray data visualization. Principal Manifolds for Data Visualization and Dimension Reduction. 2007; 58: 97-131.

[151] Gorban AN, Zinovyev AY. Principal Graphs and Manifolds. Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques, PA. 2009: 28-59. Retrieved from arXiv:0809.0490

[152] Gorban AN, Pitenko A, Zinovyev A. ViDaExpert: user-friendly tool for nonlinear visualization and analysis of multidimensional vectorial data. 2014.

[153] Gorban AN, Zinovyev AY. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. 2001.

[154] Trevor J, Hastie TRJ, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Springer; 2011.

[155] Chapman WW. A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. Journal of biomedical informatics; 2001; 34(1): 4-14.

[156] Nelder JA. A simplex method for function minimization. The computer journal; 1965; 7(4): 308-313.

[157] Price CJ. A convergent variant of the Nelder-Mead algorithm. Journal of Optimization Theory and Applications; 2002; 113(1): 5-19.

[158] Powell MJ. On Search Directions for Minimization Algorithms. Mathematical Programming; 1973; 4(1): 193âĂŞ201.

[159] Yu WC. Positive basis and a class of direct search techniques. Scientia Sinica [Zhongguo Kexue]; 1979: 53âĂŤ68.

[160] Baudin, M. (2009). Nelder mead user's manual.

[161] Luus R & Jaakola TH. Optimization by direct search and systematic reduction of the size of search region. AIChE Journal; 1973; 19(4): 760-766.

[162] Rangaiah GP. Stochastic global optimization: techniques and applications in chemical engineering. World Scientific; 2010; 2.

[163] Torczon V. On the convergence of pattern search algorithms. SIAM Journal on optimization; 1997; 7(1): 1-25.

[164] Lewis RM, Torczon V. Rank Ordering and Positive Bases in Pattern Search Algorithms. Institute for computer applications in science and engineering hampton va; 1996.

[165] Davis C. Theory of positive linear dependence. American Journal of Mathematics. 1954; 76(4):733-46.

[166] Dolan ED, Lewis RM, Torczon V. On the local convergence of pattern search. SIAM Journal on Optimization; 2003; 14(2):567-83.

[167] Bingham NH, Fry JM. Regression: Linear models in statistics. Springer Science & Business Media; 2010.

[168] Rotman JJ. Advanced modern algebra. American Mathematical Soc; 2010; 114.

[169] Gorban A, Waxman C. Neural Networks for Political Forecast. Proceedings of the 1995 World Congress On Neural Networks , A Volume in the INNS Series of Texts, Monographs, and Proceedings; 1995; 1.

[170] Wilmoth DR. Intelligence and past use of recreational drugs. Intelligence. 2012; 40(1):15–22.

[171] Jones SP, Heaven PCL, Psychosocial correlates of adolescent drug-taking behaviour. Journal of Adolescence. 1998; 21(2): 127–134.

[172] Gorban AN. Algorithms for finding duplicated attributes. Institute of Computational Modelling RAS; 1501-B00(24.05.2000); (2000): 1501-B00.

[173] Hoerl AE. Application of ridge analysis to regression problems. Chemical Engineering Progress; 2000; 58(3): 54-59.

[174] Bennett KP. Decision tree construction via linear programming. Center for Parallel Optimization, Computer Sciences Department, University of Wisconsin; 1992.

[175] Frable, WJ. Major problems in pathology. WB Saunders Co., Philadelphia; 1983.

[176] Digman JM. Higher-order factors of the Big Five. Journal of Personality and Social Psychology. 1997; 73:1246–1256.

[177] DeYoung CG, Peterson JB, Higgins DM. Higher-order factors of the Big Five predict conformity: Are there neuroses of health?. Personality and Individual Differences. 2002; 33:533–552.

[178] DeYoung CG, Peterson JB, Séguin JR, Tremblay RE. Externalizing behavior and the higher order factors of the Big Five. Journal of Abnormal Psychology. 2008; 117:947–953.

[179] Fridberg DJ, Vollmer JM, O'Donnell BF, Skosnik PD. Cannabis users differ from non-users on measures of personality and schizotypy. Psychiatry research. 2011 Mar 30;186(1):46–52.

[180] Pailing A, Boon J, Egan V. Personality, the Dark Triad and violence. Personality and Individual Differences. 2014 Sep 30;67:81-6.

[181] Whiteside SP, Lynam DR. The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. Personality and Individual Differences. 2001; 30:669–689.

[182] Gibbons S. 'Legal Highs'–novel and emerging psychoactive drugs: a chemical overview for the toxicologist. Clinical Toxicology. 2012; 50:15–24.

[183] Nutt D, King LA, Saulsbury W, Blakemore C. Development of a rational scale to assess the harm of drugs of potential misuse. The Lancet. 2007; 369:1047–1053.