STATS 503 Project Report:

Drug Use Classification by Personality Traits and Demographic Characteristics

Abstract

Drug use/misuse has become common in the twenty-first century, and identifying factors that increases an individual's risk of drug consumption has become ever important. In this report, we aim to apply four classification models to UCI Machine Learning Repository's Drug Consumption data set. Using logistic regression, k-nearest neighbors (KNN), support vector machines (SVM), and random forests, we seek to highlight the similarities and differences between these four models. Our findings indicate that different models will find similar variables more important and personality traits play an important role in classifying drug users. Nevertheless, all models found it easier to correctly classify those with heavy drug use (defined as using more than 5 different drugs in their lifetime) than those with less drug use. This perhaps implies that heavy drug users are characteristics that are more similar, making them easier to classify. Finally, we find that our logistic regression and KNN models have lower errors than our SVM and random forest models. This gives credence to Occam's Razor which states that at times simpler models are better.

1 Introduction

The evaluation of the risk for drug misuse has been a polemical topic since the 1970's War on Drugs. Drug use/misuse has many reasons: relieving pains, easing stress, or performing better at work. In this study, we aim to classify an individual's drug use based on their personality traits and demographic characteristics. We then explore how these personalities and demographic characteristics affect the use of different drugs.

1.1 Data Description

Our data comes from UCI's Machine Learning Repository. The data set contains 1885 observations with 32 variables. The variable `ID` works as the reference variable. The following 5 variables - `Age`, `Gender`, `Education`, `Country`, `Ethnicity` - are standard demographic variables. We transformed the 5 demographic variables into categorical variables based on the detailed information given by the collectors in the documentation. The next 7 variables - `Nscore` (Neuroticism), `Escore` (Extraversion), `Oscore` (Openness), `Ascore` (Agreeableness), `Impulsive`, `Cscore` (Conscientiousness), and `SS` (Sensation) - are personality measurements which, the data collectors believe, influence drug consumption. All these personality measurements have been standardized (mean 0 and standard deviation 1). The last 19 variables are categorical variables on substance use, including `Alcohol`, `Amphet` (amphetamines), `Amyl` (amyl nitrite), `Benzos` (benzodiazepine), `Caffeine`, `Cannabis`, `Chocolate`, `Coke`, `Crack`, `Ecstasy`, `Heroin`, `Ketamine`, `Legalh` (legal highs), `LSD` (lysergic acid diethylamide), `Meth` (methadone), `Mushrooms`, `Nicotine`, `Semer` (fictitious drug Semeron), and `VSA` (volatile substance abuse consumption). Each substance is categorized into 7 classes measuring when the substance was last used: Never, Over a Decade Ago, Within the Last Decade, Within the Last Year, Within the Last Month, Within the Last Week, and Within the Last Day. We transformed them into binary classes by using "Never" to form class "Never Used" and others to form class "Used".

(Detailed summary statistics of our data set is presented in Tables in Appendix.)

1.2 Variables of Interest

We used 5 demographic variables: `Age`, `Gender`, `Education`, `Country`, `Ethnicity` and 7 personality measurements: `Nscore`, `Escore`, `Oscore`, `Ascore`, `Impulsive`, `Cscore`, and `SS` as predictors. In addition, we also used 3 common substances, including `Alcohol`, `Choc` and `Caff`, to help predict.

Based on our data, we found that the average number of drugs (out of 16 drugs except `Alcohol`, `Choc` and `Caff`) a person has used is 6. We set a new variable called `MoreThan5` to show the number of a person has used is lower or higher than the average and used this variable as our response. If a person has used more than 5 drugs, then `MoreThan5 = 1`; otherwise, `MoreThan5 = 0`.

2 Explanatory Data Analysis

2.1 Numerical Predictors

2.1.1 Pair Plots

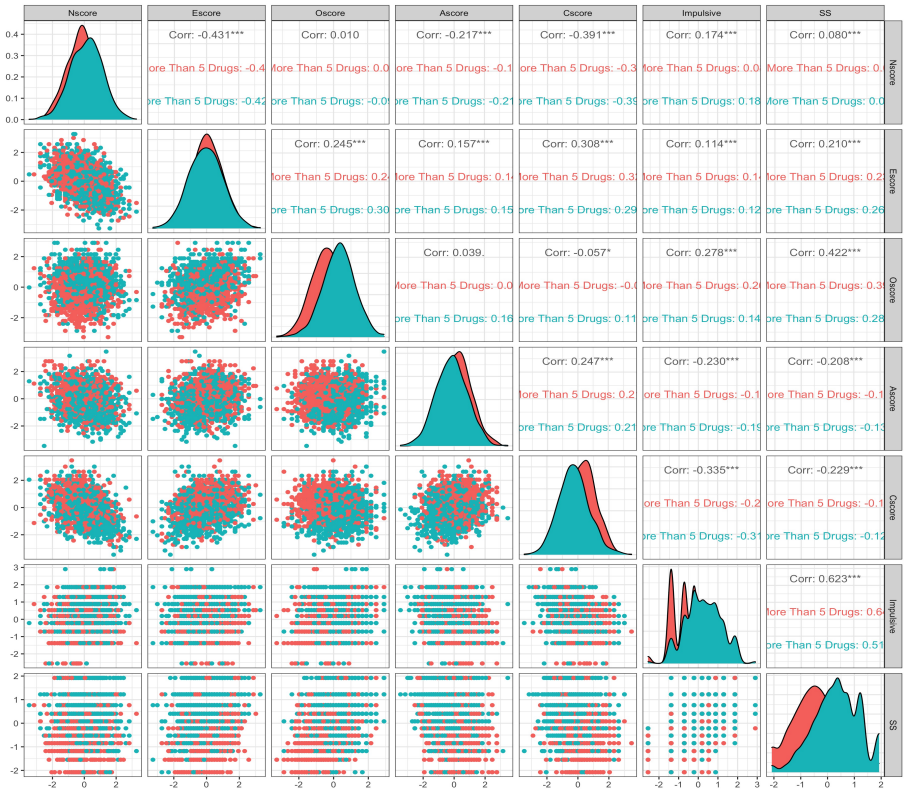


Figure 1: Pair Plots for Numerical Predictors.

According to the pair plot (Figure 1) of numerical predictors: `Nscore`, `Escore`, `Oscore`, `Ascore`, `Impulsive`, `Cscore`, and `SS`, we found that

- There is no severe collinearity problem among numerical variables;
- If only use two variables to separate `MoreThan5`, no variable-pair can do the classification very well.

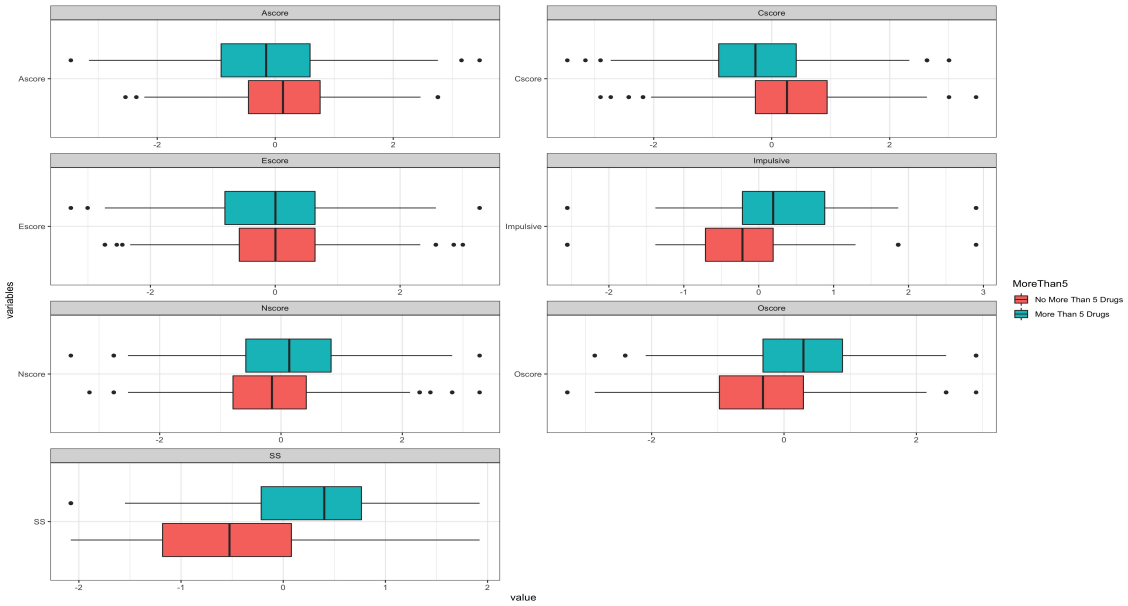


Figure 2: Box-plots for Numerical Predictors.

2.1.2 Box-plots

As the Box-plots of numerical predictors (Figure 2) show, it can be seen that people using more than 5 drugs tend to be less agreeable (lower `Ascore`), less conscientious (lower `Cscore`), more neurotic (higher `Nscore`), more open (higher `Oscore`), more impulsive (higher `Impulsive`) and more sensitive (higher `SS`).

2.2 Categorical Predictors

2.2.1 Mosaic Plots

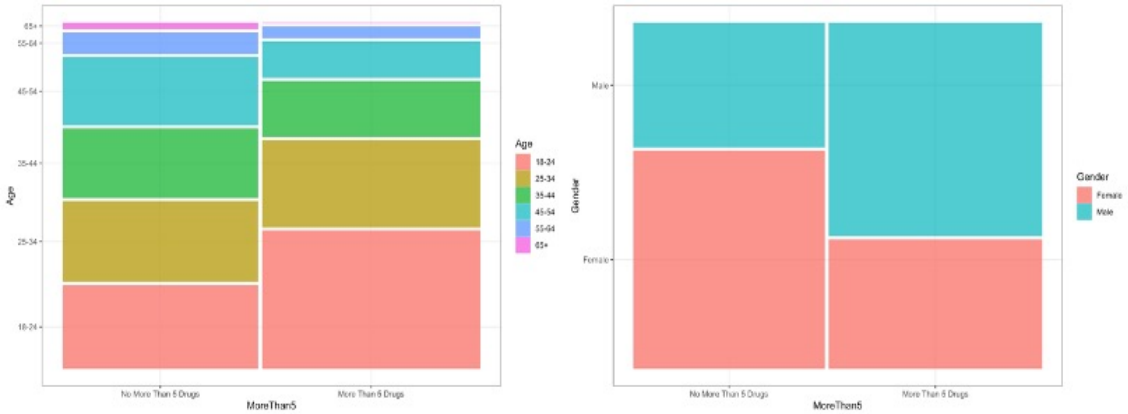


Figure 3: Mosaic Plots for Categorical Predictors.

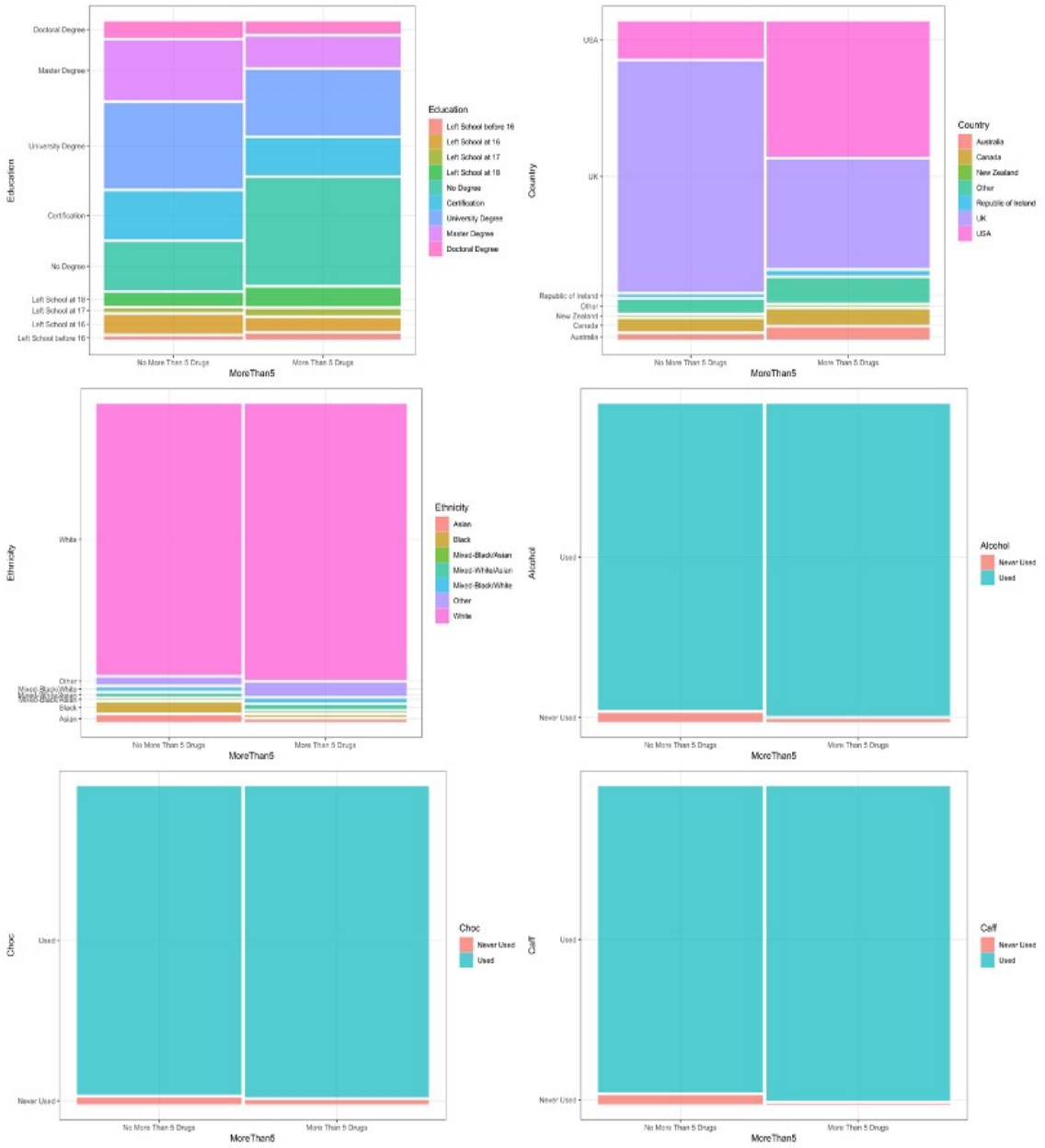


Figure 4: Mosaic Plots for Categorical Predictors.

As the mosaic plots of the categorical variables (Figure 3 and 4) show, `Age` (18-24), `Gender` (Male), `Education` (Left School at 18/No degree), `Country` (USA/ New Zealand), `Ethnicity` (White), `Alcohol` (used) and `Caff` (Used) seem to account for a higher proportion in the class "more than 5 drugs" than the proportion they account for in the class "no more than 5 drugs". And they probably can be used in the classification and have a good classification result.

3 Problem & Proposed Models

3.1 Problem Description

We learnt from the explanatory data analysis that the drug use (more than 5 drugs or no more than 5 drugs) is indeed related to the some personalities and demographic characteristics. So, we decided to explore the relationship

between the drug use and personalities, demographic characteristics further in details.

Our problem has two parts. One is on the **prediction**: we aimed to classify an individual's drug use (predict `MoreThan5`) based on their personality traits and demographic characteristics, and achieve high accuracy of predictions. The other one is on the **explanation**: we tried to find which personalities and demographic characteristics are significant in predicting the drug use (`MoreThan5`) and explore how these significant predictors affect the drug use, i.e. increasing or decreasing the risk of using more than 5 drugs.

3.2 Proposed Models

We split 80% of the whole data set as the training data and the rest 20% as the test data. Using the 5 baseline variables, 7 personality measurements and uses of 3 common substances as predictors, we attempted to classify whether one has used 5 or more different drugs, with four classification models: **k-nearest neighbors** (KNN) for its simplicity (simple to understand) and high flexibility (non-parametric), **support vector machine** (SVM) for its good performance in high dimensional space with proper choice of kernels, **logistic regression** for its interpretability (convenient to interpret model coefficients as indicators of feature importance), and **random forest** for its suitability for categorical predictors. We then selected the most effective classifier (with highest prediction accuracy) and compare the significance or importance of predictors in logistic regression and random forest models.

4 Experimental Results & Analyses

4.1 k-nearest neighbors

4.1.1 Optimal Parameters

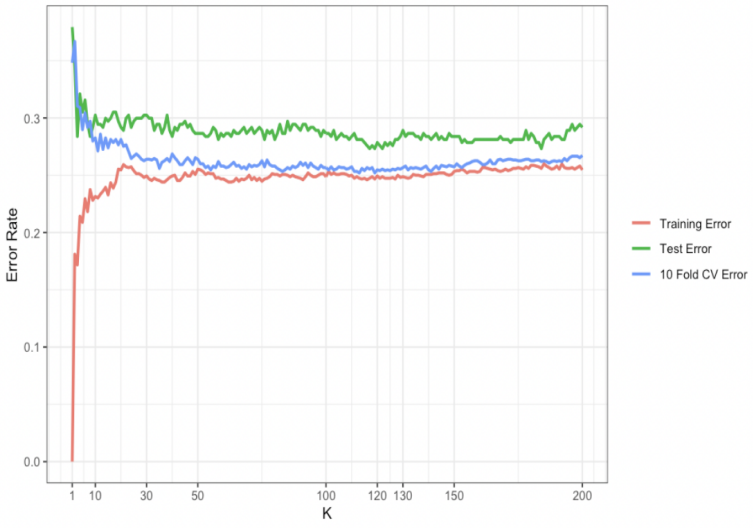


Figure 5: Error Rates vs. `K` for KNN.

By 10-fold cross validation, we found that the optimal `K` is 113, which gives a 10-fold minimum cross validation error of 0.251 (as shown in Figure 5).

4.1.2 Training & Test Errors

	Training Error	Test Error
Overall	0.247	0.281
Use no more than 5 drugs	0.304	0.384
Use more than 5 drugs	0.196	0.190

Table 1: `113`-NN Classification Errors for Each Class.

The `113`-NN model gave a test error of 0.281. The training and test errors for class "no more than 5 drugs" is higher than the class "more than 5 drugs".

4.2 Support Vector Machines

4.2.1 Optimal Parameters

In Support Vector Machines, we first cross-validated our parameters `kernel`, `cost`, `degree`, and `gamma`. The results are presented in Table 2 and we can see that the optimal parameters are `kernel = polynomial, cost = 0.5` and `degree = 2`.

Kernel	Cost	Gamma	Degree	5-fold CV Error
Linear	10	NA	NA	0.2718
Polynomial	0.5	NA	2	0.242
Radial	1	0.5	NA	0.2871

Table 2: 5-fold Cross Validation Results for SVM.

4.2.2 Training and Test Errors

	Training Error	Test Error
Overall	0.242	0.308
Use no more than 5 drugs	0.293	0.418
Use more than 5 drugs	0.147	0.210

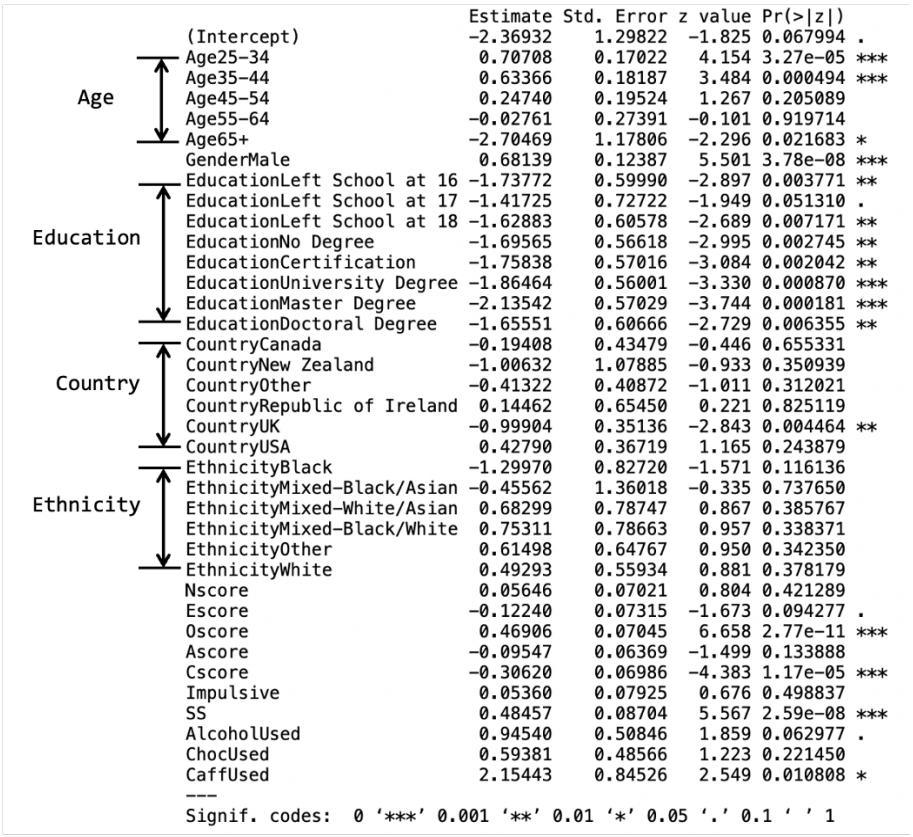
Table 3: SVM Classification Errors for Each Class.

The `polynomial`-SVM with `cost=0.5` and `degree = 2` gave a test error of 0.308. The training and test errors for class "no more than 5 drugs" is higher than the class "more than 5 drugs".

4.3 Logistic Regression

4.3.1 Coefficients of Predictors

The estimations of coefficients for each predictors are shown in Figure 6.



	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.36932	1.29822	-1.825	0.067994	.
Age25-34	0.70708	0.17022	4.154	3.27e-05	***
Age35-44	0.63366	0.18187	3.484	0.000494	***
Age45-54	0.24740	0.19524	1.267	0.205089	
Age55-64	-0.02761	0.27391	-0.101	0.919714	
Age65+	-2.70469	1.17806	-2.296	0.021683	*
GenderMale	0.68139	0.12387	5.501	3.78e-08	***
EducationLeft School at 16	-1.73772	0.59990	-2.897	0.003771	**
EducationLeft School at 17	-1.41725	0.72722	-1.949	0.051310	.
EducationLeft School at 18	-1.62883	0.60578	-2.689	0.007171	**
EducationNo Degree	-1.69565	0.56618	-2.995	0.002745	**
EducationCertification	-1.75838	0.57016	-3.084	0.002042	**
EducationUniversity Degree	-1.86464	0.56001	-3.330	0.000870	***
EducationMaster Degree	-2.13542	0.57029	-3.744	0.000181	***
EducationDoctoral Degree	-1.65551	0.60666	-2.729	0.006355	**
CountryCanada	-0.19408	0.43479	-0.446	0.655331	
CountryNew Zealand	-1.00632	1.07885	-0.933	0.350939	
CountryOther	-0.41322	0.40872	-1.011	0.312021	
CountryRepublic of Ireland	0.14462	0.65450	0.221	0.825119	
CountryUK	-0.99904	0.35136	-2.843	0.004464	**
CountryUSA	0.42790	0.36719	1.165	0.243879	
EthnicityBlack	-1.29970	0.82720	-1.571	0.116136	
EthnicityMixed-Black/Asian	-0.45562	1.36018	-0.335	0.737650	
EthnicityMixed-White/Asian	0.68299	0.78747	0.867	0.385767	
EthnicityMixed-Black/White	0.75311	0.78663	0.957	0.338371	
EthnicityOther	0.61498	0.64767	0.950	0.342350	
EthnicityWhite	0.49293	0.55934	0.881	0.378179	
Nscore	0.05646	0.07021	0.804	0.421289	
Escore	-0.12240	0.07315	-1.673	0.094277	.
Oscore	0.46906	0.07045	6.658	2.77e-11	***
Ascore	-0.09547	0.06369	-1.499	0.133888	
Cscore	-0.30620	0.06986	-4.383	1.17e-05	***
Impulsive	0.05360	0.07925	0.676	0.498837	
SS	0.48457	0.08704	5.567	2.59e-08	***
AlcoholUsed	0.94540	0.50846	1.859	0.062977	.
ChocUsed	0.59381	0.48566	1.223	0.221450	
CaffUsed	2.15443	0.84526	2.549	0.010808	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Figure 6: Logistic Regression Results.

It can be found that the predictors `Age`, `Gender`, `Education`, `Country`, `Escore`, `Oscore`, `Cscore`, `SS`, `Caff` and `Alcohol` are significant under the 0.1 level of significance.

Analysis of Significant Predictors:

- The coefficient of `GenderMale` is estimated as 0.6, which indicates that, compared with Female, `Male` are more likely to use more than 5 drugs;
- The coefficients of `Escore`, `Oscore`, `SS` and `Cscore` are estimated as -0.12, 0.44, 0.48 and -0.30 separately. Therefore, we can say that lower Escore (less extraversive),higher Oscore (more open), higher SS (more sensitive) and lower Cscore (less conscientious) will increase the probability of using more than 5 drugs;
- Both the estimations of coefficients of `AlcoholUsed` and `CaffUsed` are positive, which indicates having used Alcohol and Caffeine will increase the probability of using more than 5 drugs.
- `Age`
  - The reference level of `Age` is `18-24`;
  - Compared with `18-24`, `25-34`, `35-44` and `45-54`, whose coefficients are estimated as positive, are more likely to use more than 5 drugs and `55-64` and `65+`, whose coefficients are estimated as negative, are less likely to use more than 5 drugs;
  - Based on the value of coefficients, we may say that the probability of using more than 5 drugs: `25-34` $>$ `35-44` $>$ `45-54` $>$ `18-24` $>$ `55-64` $>$ `65+`;
- `Education`
  - The reference level of `Education` is `Left School before 16`;

- All coefficients of other levels of Education are estimated as positive, so all the other levels can decrease the probability of using more than 5 drugs;
- The probability of using more than 5 drugs: `Left School before 16` > `Left School at 17` > `Doctoral Degree` > `Certification` > `Left School at 16` > `Left School at 18` > `No Degree` > `University Degree` > `Master Degree`;

- `Country`
  - The reference level is `Australia`;
  - Compared with Australia, UK decreases the probability of using more than 5 drugs significantly;
  - Probability of using more than 5 drugs: `USA` > `Republic of Ireland` > `Australia` > `Canada` > `Canada` > `New Zealand` > `New Zealand`;

4.3.2 Training and Test Errors

	Training Error	Test Error
Overall	0.235	0.244
Used no more than 5 drugs	0.261	0.282
Used more than 5 drugs	0.218	0.210

Table 4: Logistic Regression Classification Errors for Each Class.

The logistic regression model gave a test error of 0.244. The training and test errors for class "no more than 5 drugs" is higher than the class "more than 5 drugs".

4.4 Random Forests

4.4.1 Optimal Parameters

In random forest, averaged 'out-of-bag' (OOB) errors are used to estimate the test error and select the optimal the number of variables randomly sampled as candidates at each split (`mtry`) and minimum size of terminal nodes (`nodesize`) in random forest model.

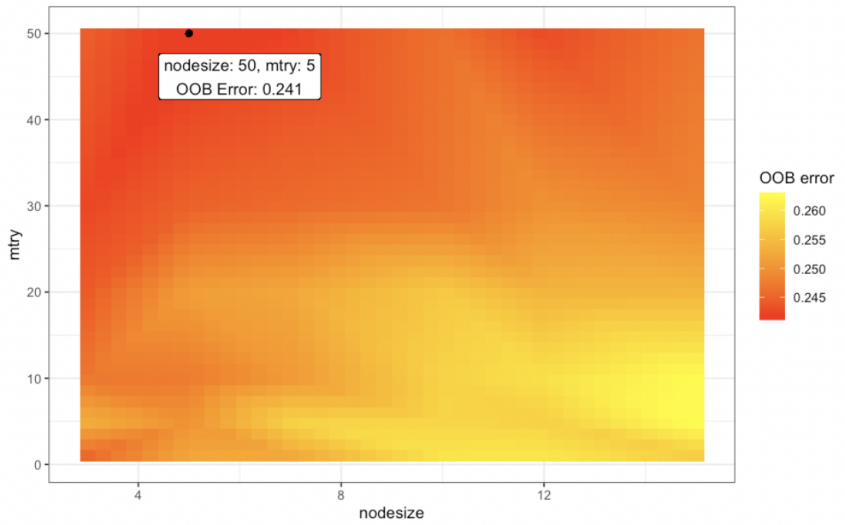


Figure 7: OOB Error for Different `nodesize` & `mtry`.

As the Figure 7 shows, we found that `mtry = 5` and `nodesize = 50` achieves the minimum OOB error at 0.241.

4.4.2 Importance Rating

Figure 8 shows that the first 8 important predictors given by mean decrease of Gini are `Country`, `SS`, `Oscore`, `Cscore` `Impulsive` and `Nscore`, `Ascore` and `Education`. Among these 8 predictors, the most important predictor is `country`, which indicates that the use of drugs may be influenced by different laws or cultures of different countries; moreover, there are 6 personality variables among 8, and we may conclude that the personality is pretty important for predicting the use of drugs.

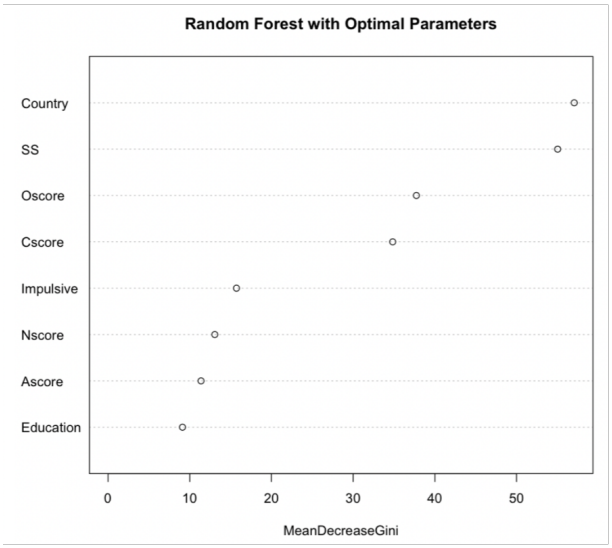


Figure 8: Importance Plot for Predictors Used in Random Forest with Optimal Parameters.

4.4.3 Training and Test Error

The random forest model with the optimal parameters then provides a training error of 0.198 and a test error of 0.289. Classification error rates for each class is shown in Table 5.

	Training Error	Test Error
Overall	0.198	0.289
Used no more than 5 drugs	0.266	0.373
Used more than 5 drugs	0.136	0.215

Table 5: Random Forest Classification Error for Each Class.

The random forest model gave a test error of 0.289. The training and test errors for class "no more than 5 drugs" is higher than the class "more than 5 drugs".

5 Conclusion

5.1 Prediction

The classification errors of four models: KNN, SVM, Logistic Regression and Random Forest with optimal parameters in section 4 are shown as Table 6.

And it can be concluded that

- In terms of overall test error, Logistic regression perform best; This is a strong reminder that at times "simple" models can outperform more sophisticated models as implied by Occam's Razor.
- In terms of test error for the class "use more than 5 drugs", KNN performs best.
- In terms of test error for the class "use no more than 5 drugs", no model performs very well. We think one possible reason why models perform better for "use more than 5 drugs" than "use no more than 5 drugs" can be that people heavily using drugs probably have more common characteristics, but people using less drugs have various background and characteristics.

Model	KNN	SVM	Logistic Regression	Random Forest
Overall Training Error	0.247	0.242	0.235	0.198
Overall Test Error	0.281	0.308	0.244	0.289
Training Error No More Than 5 Drugs	0.304	0.293	0.261	0.266
Test Error No More Than 5 Drugs	0.384	0.418	0.282	0.373
Training Error More Than 5 Drugs	0.196	0.147	0.218	0.136
Test Error More Than 5 Drugs	0.190	0.210	0.210	0.215

Table 6: Classification Errors for Overall and Each Class.

5.2 Explanation

- Based on random forest importance rating, `Country, SS, Oscore, Cscore, Impulsive` and `Nscore` are important to classify people using more than 5 drugs and using no more than 5 drugs; And personality is more important for predicting the use of drugs than demographic variables.
- Based on logistic regression, `Gender, Education, Country, Oscore, Cscore, SS, Caff` and `Alcohol` are significant to classify, which are similar to the important predictors given by random forest; Besides, male, more open, more sensitive, less conscientious, having used alcohol and caffeine, left school before 16 and from USA are more likely to use more than 5 drugs.

5.3 Limitations

- There are still some potential aspects that could be explored in this problem. Though we just use a response representing whether one has used 5 or more kinds of drugs, we can also predict the use of each kind of drugs with the data, which could has a quite different results, and maybe we can find out some differences between different drugs.
- In this report, 4 models have been used to the problem, but there are also some other classification models that could be applied to our data set. Also, all models we used are supervised learning methods, and there exist a limitation of a lack of unsupervised learning methods like K-Means.

(All codes are attached in Appendix.)