A Data Study on Drug Use

Yuting Lu

Department of Mathematics, Imperial College London

Introduction

In this project, a drug use dataset is analyzed in order to find connections between certain characteristics of drug users like age, gender, country, ethnicity, and quantitative variables like personalty measurements and the regularity of drug use.

Exploratory data analysis

- Some visualizations that illustrate the relationship between substance use and the predictor variables.

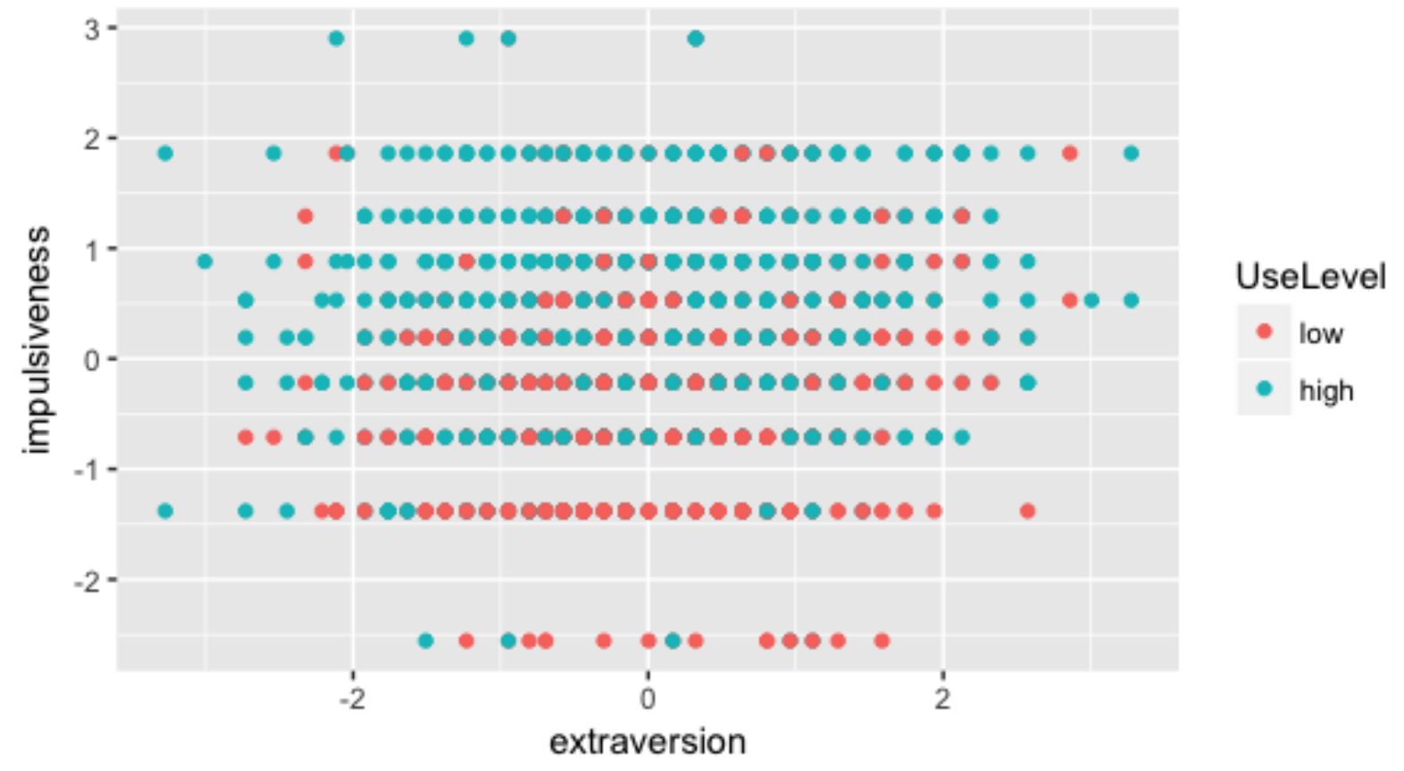

Figure 1: a scatter plot of extroversion and impulsiveness

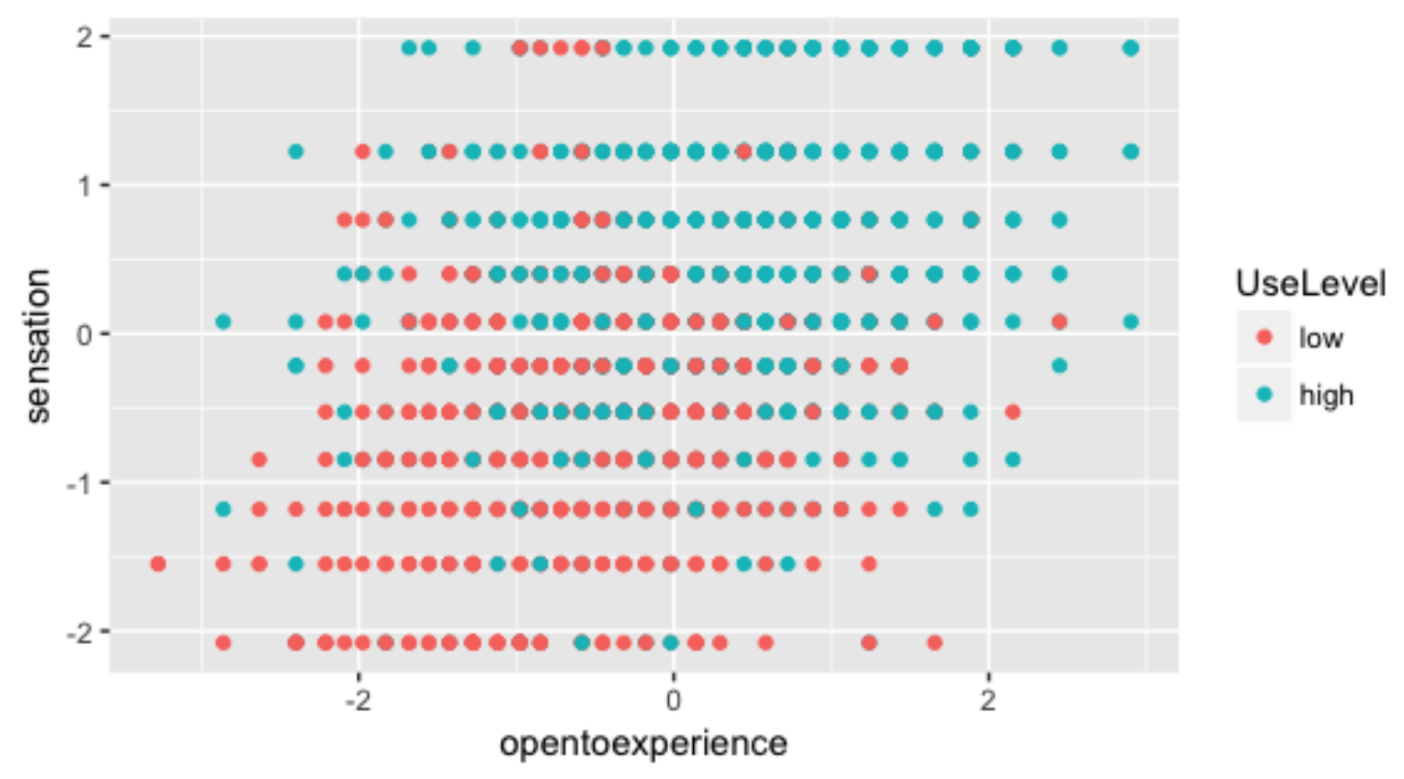

Figure 2: a scatter plot of opentoexperience and sensation

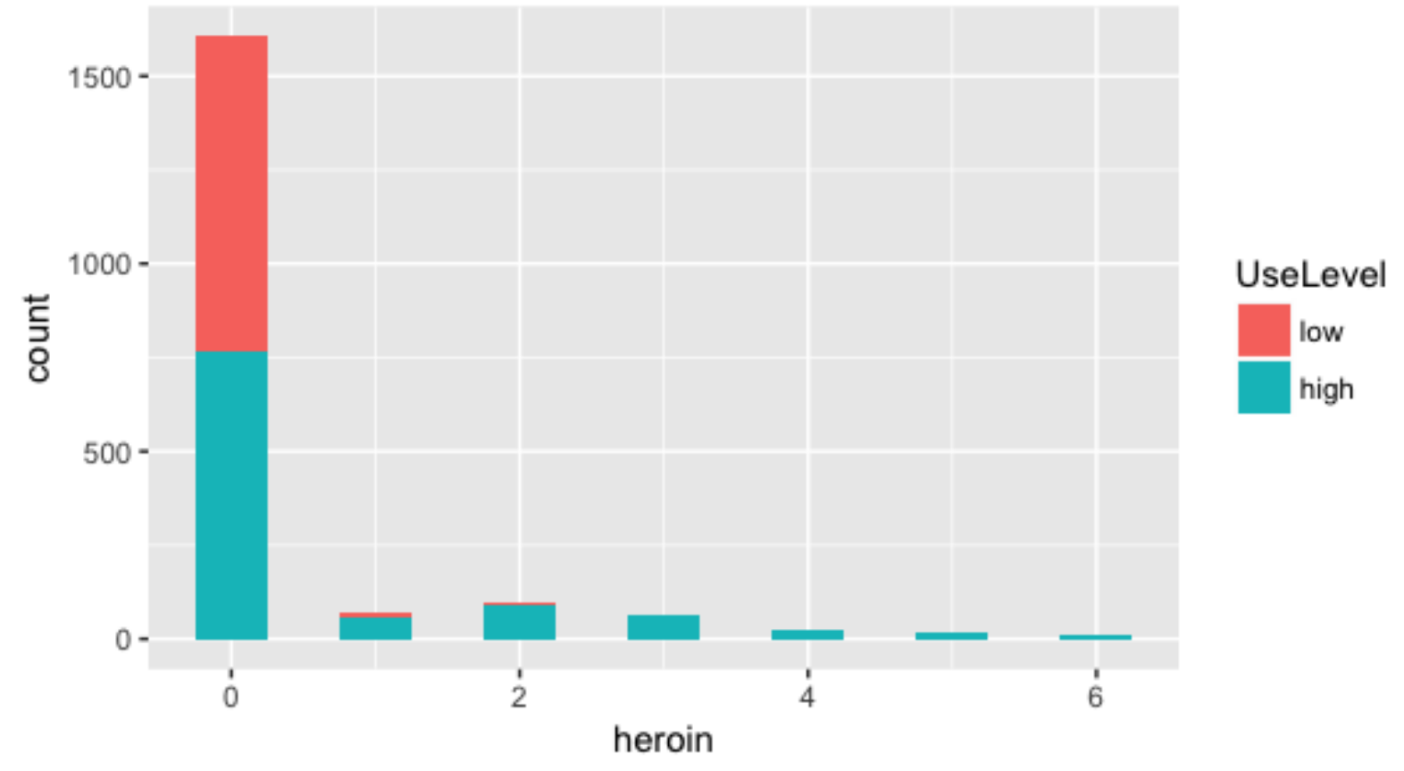

Figure 3: a bar plot of heroin use with fill=UseLevel

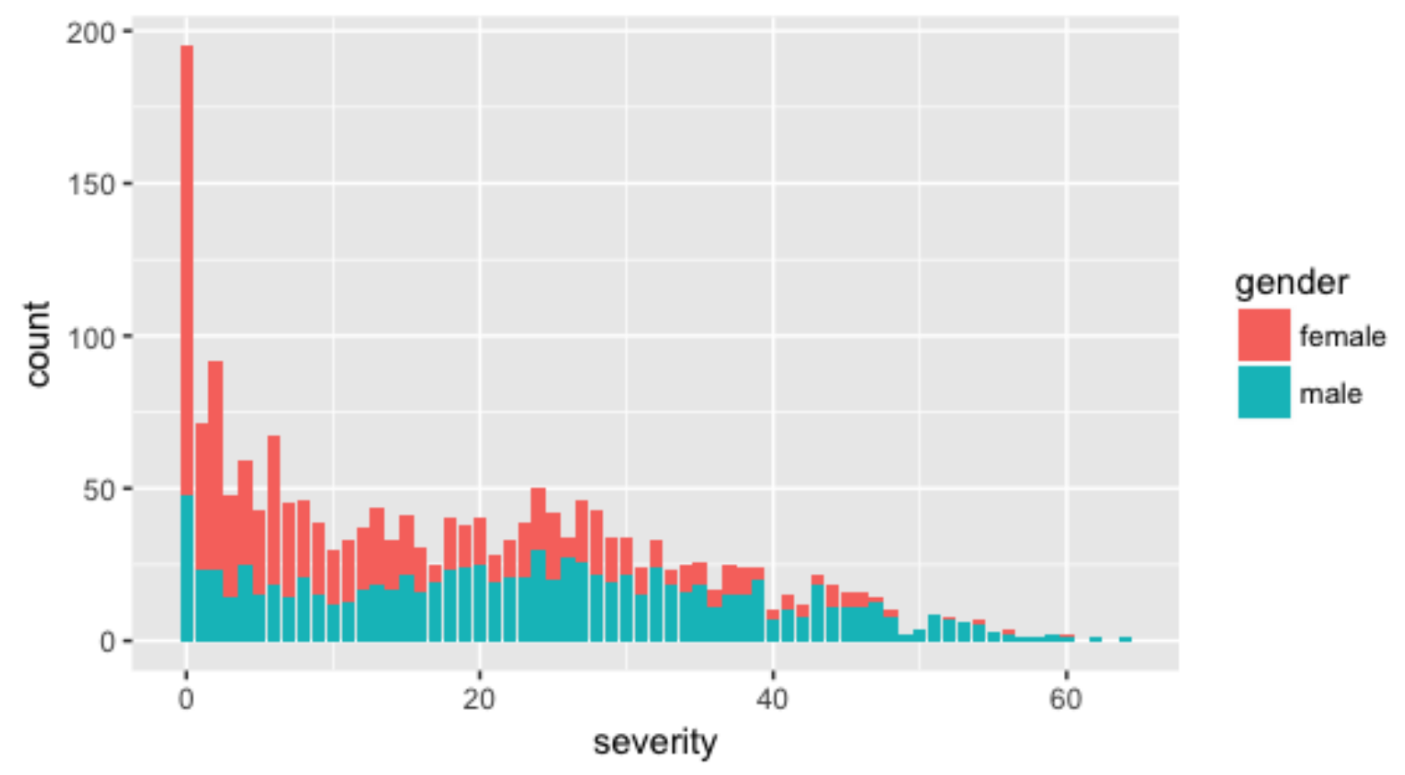

Figure 4: a bar plot of severity and gender

Observations from the plot

- Impulsiveness and high level of drug use are positively correlated. This also applies to other parameters like openness to new experiences and Sensation
- Heroin use is positively correlated to heavy substance abuse.
- The data suggests males tend to have a higher drug intake than females.

Predicting UseLevel

- Method: k-nearest neighbors algorithm.
- Predictors: personality measurements, selected categorical variables.
- Reason for this: with KNN the predictors can only be numerical, so I have to convert all the categorical variables to many columns of 0 and 1. However, the KNN might fail if there are lot of predictors and many do not relate to outcome.
- Estimation the performance by doing: 1. Adding a 10-fold cross validation. 2. Plotting a ROC curve and calculate its AUC. 3. By trial and fail, I found out that k=7 gives the highest accuracy.

Outcomes and Conclusion

Table 1: Confusion Matrix

knnpredictions32	0	1
0	76	6
1	7	99

```
[1] "Accuracy: 0.902127659574468"
[1] "AUC: 0.929259896729776"
```

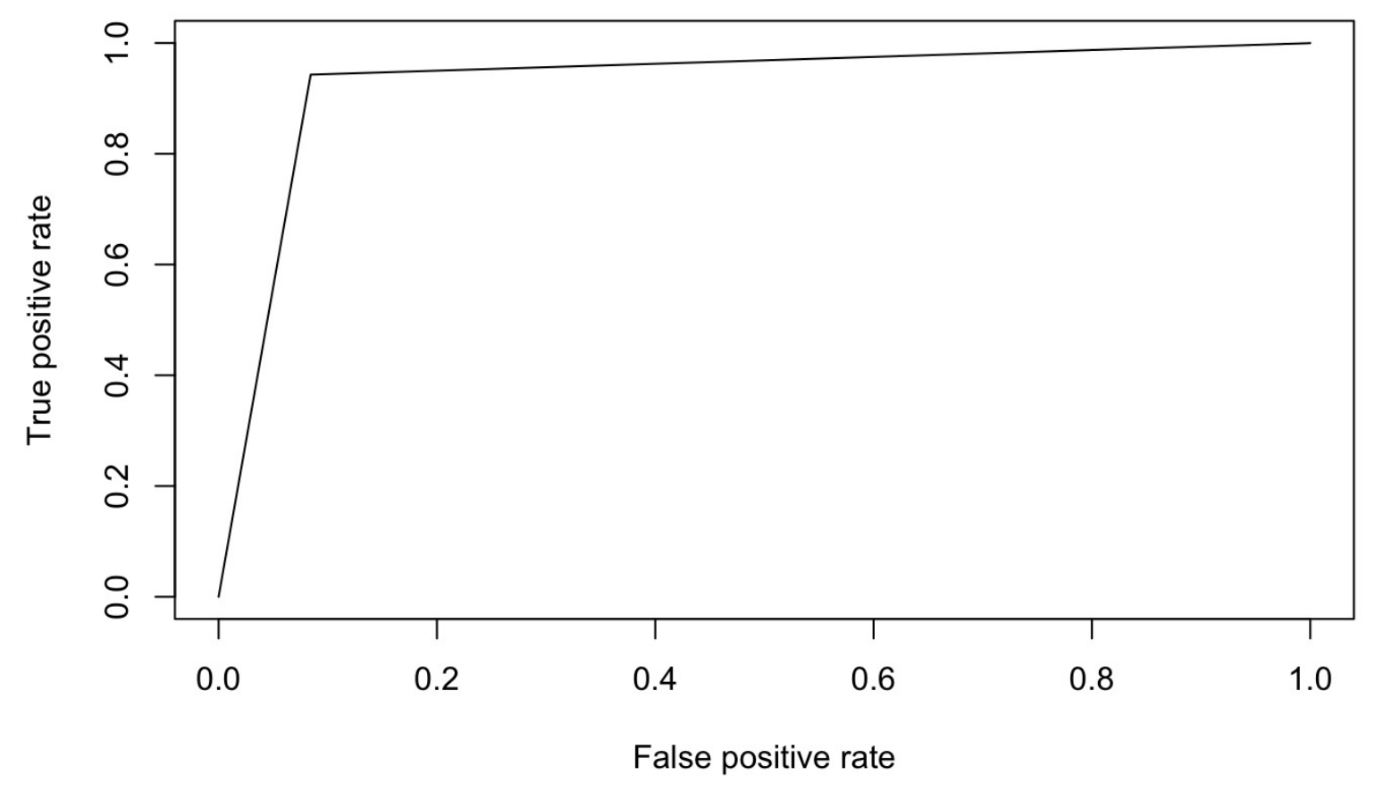

Figure 5: ROC curve for this classifier

- The accuracy of the results is high.
- The ROC curve shows a good classification result because the error is reasonably small and the area under the curve is close to 1.

Exploring the relationship between LSD use and personalities

- Question: How do personality measurements affect whether people takes LSD.
- Initial idea: LSD is not a drug for social purpose like cannabis. I am curious about the personality traits of people who use it.
- Method used: Decision tree.
- Reason :The decision tree method gives an interpretable picture of the relationships between the predictors and the classification.
- Predictors: Personality measurements and genders.
- To avoid overfitting:
  -Using bootstrap to sample the data.
  -Finding a good complexity parameter by plotting how the cross validated error rate related to various complexity parameter thresholds. Then use the complexity parameter with the lowest error to prune the tree.

Outcomes and Conclusion

Table 2: Confusion Matrix

dtreepredictions	0	1
0	308	116
1	99	181

```
[1] "Accuracy: 0.694602272727273"
```

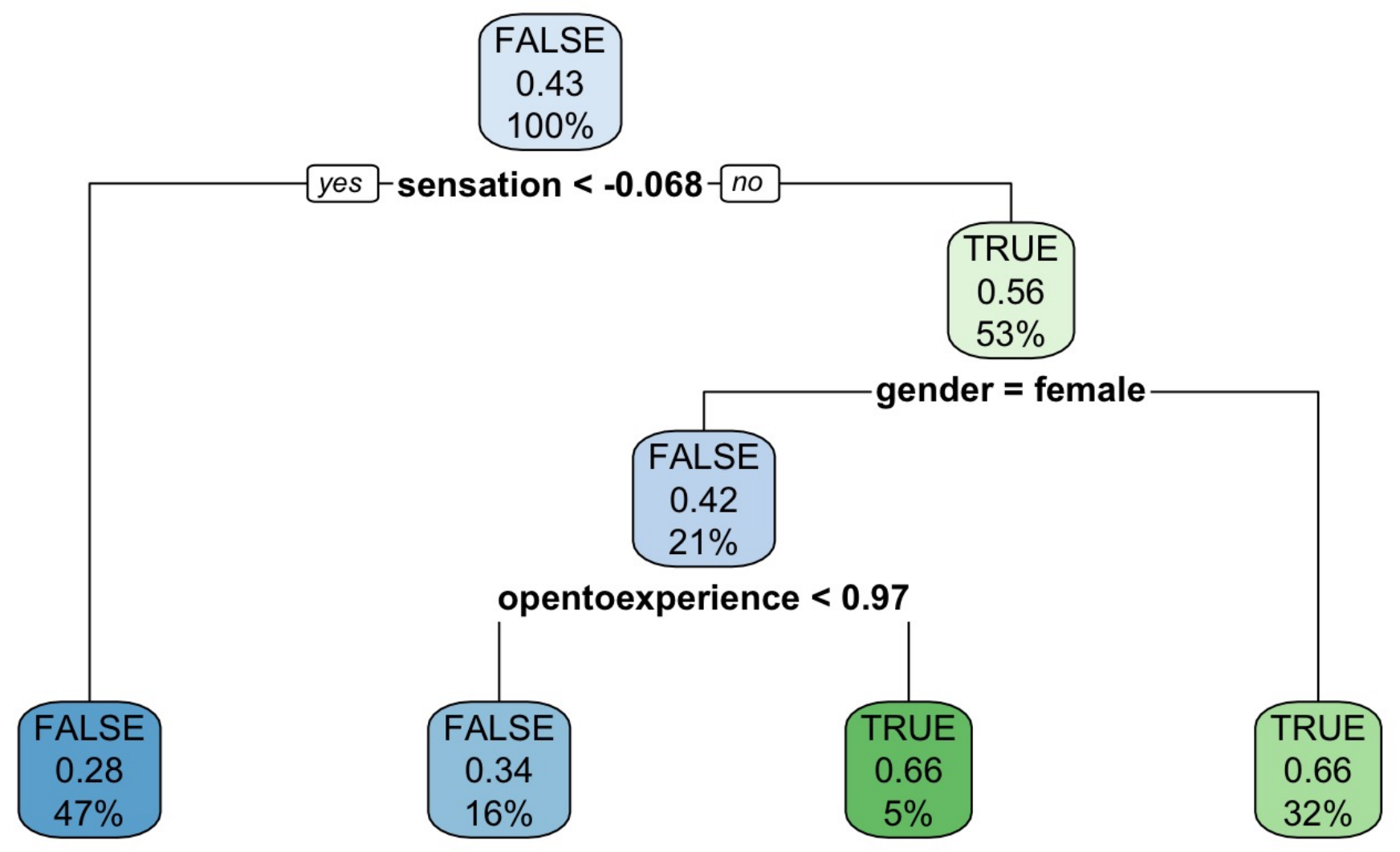

Figure 6: A pruned decision tree

- Interpretation: Each node shows
  - the predicted class :FALSE(never used) or TRUE(have used),
  - the predicted probability of using LSD,
  - the percentage of observations in the node.
- Conclusion: People who have higher sensation, are more open to experience tend to use LSD. In addition, the data suggest LSD use is higher amongst males.

Further discussion

- From the EDA, personality measurements like impulsiveness, sensation openness to experience and gender are important predictors for the level of drug use of a person. In addition, the use of illegal drugs seems to be associated with more severe intakes.
- From the logistic regression, we are told that the significant predictors are agegroup 45-54, 55-64, gender male, education, country Other, country Republic of Ireland, Country UK, open to experience, conscientiousness, impulsiveness, nicotine. They contributes to the model more than the other predictors and therefore, are important.
- After including the illegal drugs, with the help of "varImpPlot" it can be seen that the use of other illegal drugs like crack and methadone is important in predicting heroin use as shown in the picture.

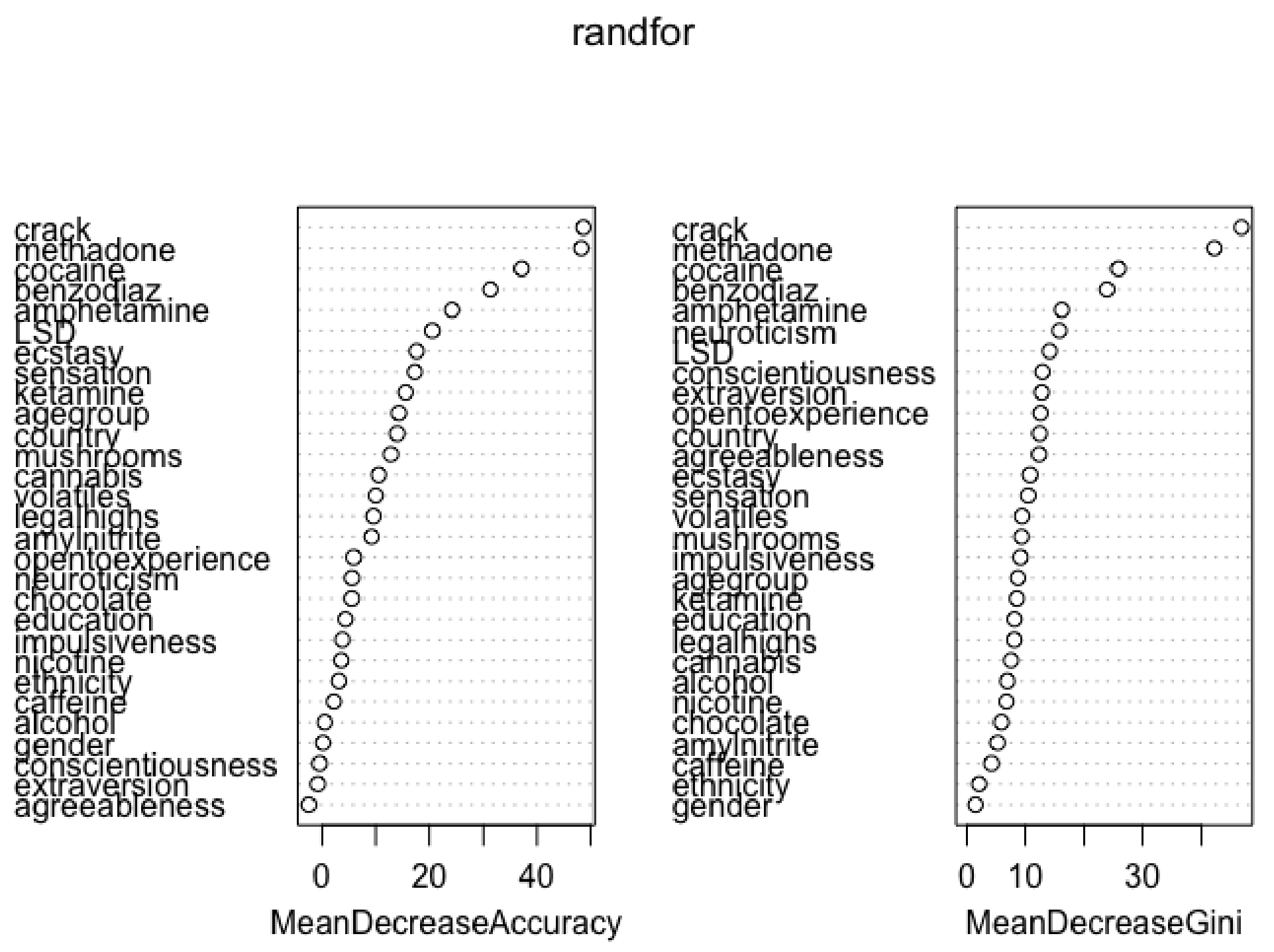

Figure 7: Plot of important variables in random forest

The contents of this work and the associated code are my own unless otherwise stated.