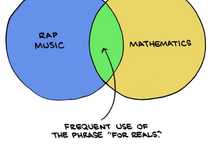


How can Personal Data Predict Drug Use?

Christopher R. McLeod

Imperial College London

Introduction

In this project, we have a dataset of 1885 data points categorised into a mix of 34 continuous & categorical measurements and drug use readings. Some notable measurements include

- categorical: age group, ethnicity, UseLevel ...
- numerical: opentoexperience, impulsiveness

In this poster, we will see exploratory analysis of the data, a K-Nearest-Neigbours classification for UseLevel and a random forest to predict cannibis use.

Credit and thanks for the poster template goes to Emma McCoy.

Exploratory Analysis

Below are some exploratory plots comparing different measurements

Predicting Cannabis use by Drug Use & Personality Metrics

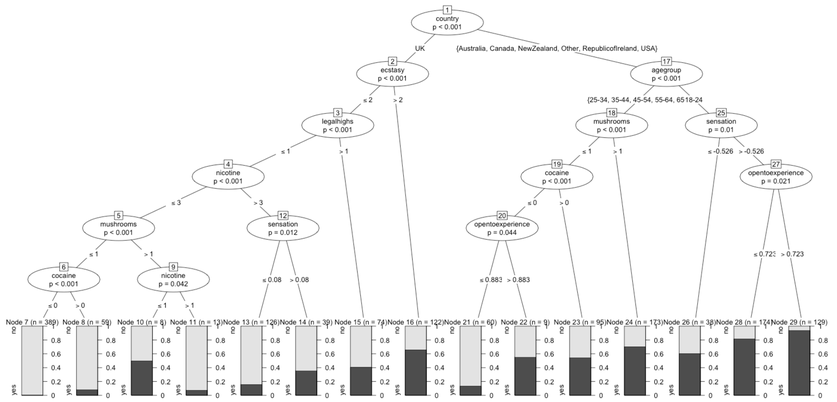
Pose the following questions:
- **Q1**: what information do we need to know about someone to predict their propensity to use cannabis regularly?
- **Q2**: Does one's country affect the likelihood one uses cannabis regularly?

Method : Train a random forest to sort for best predictors - is country significant?
- Data in : 16 personality predictors, drug information
- Response : "yes" - regular cannabis use; "no" - otherwise

```
RFcan <- randomForest(cannabis  .  , data=can.train, importance = TRUE, ntree = 2000)
can.prediction <- predict(RFcan, can.test)
```

Example Tree

Here we see the visualisation of an example tree



- Method ran once
- Importance function called
- Method ran again for 10 most significant predictors

Results

	MeanDecreaseAccuracy
agegroup	58.37245
gender	26.39574
country	104.30951
opentoexperience	32.24647
sensation	47.30529
nicotine	53.36737
cocaine	39.95916
ecstasy	61.13762
legalhighs	58.79608
mushrooms	69.93532

Bigger MDA means larger significance to random forest prediction

Answers

First run: Accuracy = 0.8381
Second run: Accuracy = 0.8170
Conclusion: Model runs well for sparser predictions

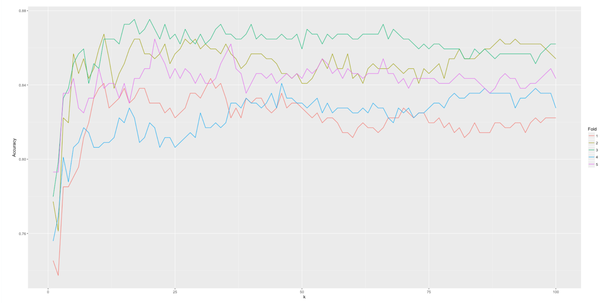
A1: The 10 in the results section are most likely a good starting point

A2: One's country appears to be a very significant predictor for regular cannabis use

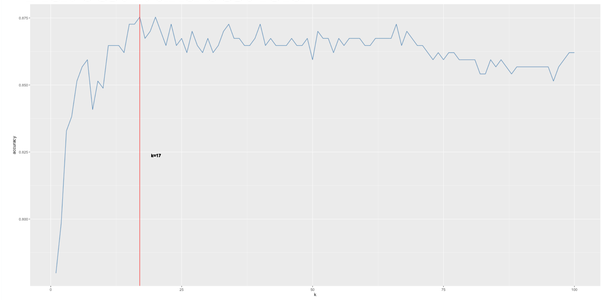
Using KNN to Predict UseLevel

Q: Predict UseLevel using first 16 personality predictors
Method: Use KNN. Perform 5-cross-fold-analysis and vary 'k' parameter at the same time. Pick best fold & best k to predict

```
predictions <- knn(train.nk[, -1], test.nk[, -1], train.nk$UseLevel, k = k)
```



- Pick best k and best fold



Results: Accuracy = 0.8753

prediction	low	high
low	151	23
high	24	179

Significant Predictors

Exploratory results and linear regression show the following are significant predictors for drug use
- Gender
- Opentoexperience
- Sensation
- Nicotine
- Country
- agegroup (potentially)