Machine Learning Prediction of Drug Overdoses among Young People

Luke Liangzhi Dai-Woodys, PhD
University of Dundee

Contact
linkedin.com/in/luke-dai-woodys

HTA343

Objectives

Drug overdose deaths are rising among young people in the United States.

This study examines whether machine learning can predict drug overdoses among young people.

Data

Datasets

- United States National Survey on Drug Use and Health from 2005 to 2014

Sample

- 356704 observations
- Age: 15-24 years

Modelling

Models

- Classification: Decision Tree
- Regression: Logistical Regression

Variable categories

- Demographics, drug use, drug dependence, drug abuse, mental health, insurance status, health behaviours and health conditions

Different sizes of sample

- 10%, 30%, 50% and 80% proportions of the sample
- To understand if a smaller sample size can predict drug overdoses as well as the full size

Information gains of variables

- Variables with high information gains can indicate a simplified portfolio for further applications.

Results

Area under the curve

- Decision Tree yields an area under the curve (AUC) of 0.66 with 95% confidential intervals (CI) between 0.62 and 0.71.
- Logistical Regression shows an AUC of 0.62 (95% CI: 0.58-0.67).
- AUCs of the two models are over 0.5, indicating the effectiveness of both models on overdose prediction.
- The AUC of Decision Tree is larger than the AUC of Logistical Regression, implying a better performance of classification than regression.

Table 2 Measures for model performance

	Decision Tree	Logistical Regression
AUC	0.66	0.62
95% CI	[0.62, 0.71]	[0.58, 0.67]
Sensitivity	0.07	0.52
Specificity	0.94	0.64

Different sample sizes

- With the increasing proportions of training sets, the AUCs of Decision Tree and Logistical Regression stay around 0.65 and 0.62 respectively.
- The AUCs of above proportions of training sets cover the AUCs of the full set.

Table 2 AUCs under different sizes of training sets

	Decision Tree	Logistical Regression
10%	0.63	0.62
30%	0.65	0.61
50%	0.66	0.62
80%	0.65	0.62

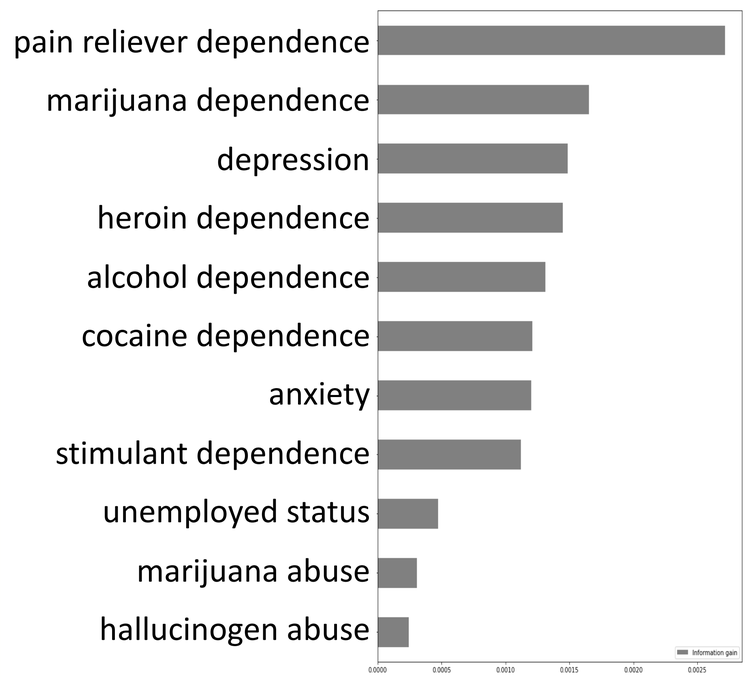
Results

Information gains

Variables with high information gains:

pain reliever dependence, marijuana dependence, depression, heroin dependence, alcohol dependence, cocaine dependence, anxiety, stimulant dependence, unemployed status, marijuana abuse and hallucinogen abuse.

Figure 1 Variables with high gains



Conclusions

- Drug overdoses among young people are predictable by machine learning.
- Classification models perform better than regression models.
- The full sample reaches the optimum on predicting drug overdoses among young people.
- Variables with high information gains are recommended for practical applications on the prediction.