Data Analysis – Drug Consumption Dataset

(Naïve Baye's Classification)

Omkar Vaidya
Computer Science
Rochester Institute of Technology
ov5232@rit.edu

Description :

The data mining component of the project is based on the Drug Consumption dataset. Performing the analysis of the dataset will help to gain insight on the factors which can influence a person to consume drugs. There are different factors involved in the consumption of drug by an individual and can vary from social, physical, geographical etc.

The dataset contains records for 1885 respondents. For each respondent 12 attributes are known : personality measurements which include (neuroticism, extraversion, openness to experience, agreeableness, conscientiousness, impulsivity, sensation seeking) , age, gender, level of education, country of residence and ethnicity. In addition participants were questioned concerning their usage of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepine, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, mushrooms, nicotine and volatile substance abuse and one fictitious drug (Semeron)). For each drug they have to select one of the answers :never used the drug, used it over a decade ago, or in the last decade, year, month, week, or day.

Motivation:

Consumption of unregulated/illegal drugs is a worldwide issue. The effects of drugs are not limited to health of the individual consuming them but also affects the people around him. The motivation behind the data mining of drug consumption data is to better understand what factors compels a person to consume drugs. The understanding of the factors will help to fight the war against the drugs. This could help to fight drugs from a much deeper level compared to just banning them. Governments could use the result of analysis of this data to understand the factors which influence the drug usage and try to implement measures to minimize drug use. Thus, data mining of this data set will prove beneficial for betterment of society.

Data Mining Technique :

The data mining of the dataset will be done by performing Classification of the dataset. This technique is most suited for the data set as, it provides the drug consumption patterns for individuals and we could analyze this data and with the help of the results predict drug consumption for an individual using the important attributes like Age, Gender, Education, Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness, Impulsiveness and Sensation, which are factual attributes and therefore result can be obtained for consumption of a particular drug.

Naïve Bayes Classifier is used to perform the classification. This technique is best suited for the classification of the dataset as all the attributes which determine the drug consumption are independent of each other and the drug consumption is dependent on the probabilities of individual attributes. (Age, Gender, Education, Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness, Impulsiveness and Sensation).

The computation for classification can be shown as :

P(Age, Gender, ..., SS | Drug) =

P(Age | Drug) P(Gender | Drug) .... P(SS | Drug) here Drug = Alcohol, Amphet, ... , VSA

We compute -

P(Attribute$_i$ | Drug$_j$) for all Attribute$_i$ and Drug$_j$ for the training data. Drug$_j$ is the class label.

The new point is classified to Drug$_j$ if

P(Drug$_j$) π P(Attribute$_i$ | Drug$_j$) is maximal.

Challenges :

Decoding the dataset : The entire dataset is present in the form of encoded values. The first task towards data mining was to decode this data. This was done using available conversion rules and writing R script to apply the rules and generate readable data. The task was performed using the 'revalue' function from the 'library(plyr)' in R language.

Deciding influencing attributes: In order to make the classification for the consumption of the drug, it is very important to decide the appropriate important attributes that have a significant effect on the result. This was performed by understanding and analyzing the density plots of the attributes. The input attributes chosen to the classifier are Age, Gender, Education, Neuroticism, Extraversion, Openness, Agreeableness, Conscientiousness, Impulsiveness and Sensation.

Scaling data: The numerical attributes which influence the classification are available in form of decimal numbers, consisting of both positive and negative values. This posed a problem while implementing Naive Bayes classifier in R. Thus the data set is scaled by a constant factor (k=4) (after finding the lowest negative value (-3.4) among all numerical attributes).

Conclusions from discovered knowledge :

Naive Bayes classifier is used to perform the classification of the drug consumption dataset.

Cross-Validation : The goal of performing cross-validation was to assess and obtain an insight on how our model will generalize to an independent dataset. In order to perform this, we partitioned 70% of the dataset for training purpose and 30% of the dataset for testing. The accuracy of prediction to classify the consumption of drug based on training and testing data for each drug is as follows : (All values in percentage.)

DRUG	**Train Accuracy**	**Test Accuracy**
Alcohol	40.6	39.5
Amyl	70.3	66.7
Amphet	50.9	54.0
Benzos	52.7	54.0
Coke	55.2	54.7
Caff	72.9	74.9
Cannabis	24.9	23.8
Choc	43.5	41.2
Crack	55.2	87
Ecstasy	53.9	54.9
Heroin	84.9	85.8
Ketamine	78.3	81.0
Legalh	57.3	59.8
LSD	56.6	57.0
Meth	75.0	77.8
Mushrooms	52.9	50.1
Nicotine	32.6	31.8
Semer	99.6	99.6
VSA	76.9	77.8

From the above table we can observe that there are certain drugs for which the prediction rate is high and some for which the prediction rate is less than 50%. The possible reason behind obtaining less prediction rate is one of the drawbacks of Naive Bayes classifier - If one of the conditional probabilities is zero, then the entire expression becomes zero.

Density plot visualization : From the density plots of each attribute we observed that higher the sores (NScore, EScore, OScore, AScore, CScore) greater was the recent (Used in Last Day) consumption of drug.

Lessons Learned:

Choosing data mining technique: There is a big list of data mining techniques, but to get maximum value out of the data, choice of appropriate data mining technique is required. The data mining project has helped with understanding real life implementations of data mining and helped to learn how to choose a data mining technique.

Deciding training data and testing data: The Naïve Bayes classifier works best when maximum data is used for training, but testing is also important to check the classification accuracy. The trick is to split available data into a proportion which will increase the accuracy and still have enough data left for making sure the classifier works with acceptable accuracy.

Activities Performed :

1. Operations on Dataset – Decoding the data values, Initial visualization to obtain insight on attributes of data.
2. Understand each attribute and it's values.
3. Transformation of the dataset – Scaling of dataset, Cross-validation of data (Data Partitioning)
4. Implementing Naïve Baye's classifier on the data set and prepare visual aids for the results of data mining.