Is there any relationship between drug usage and social states?

Thilo Ritzerfeld (2462966)
Moritz Richter (2620731)

Fontys Hogeschool Techniek en Logistiek
Business Informatics & Software Engineering
Data Mining (DaMi)

Venlo, December 13, 2017

Information

Title	Is there any relationship between drug usage and social states?
Name of the authors	Moritz Richter (2620731)
	Thilo Ritzerfeld (2462966)
Module	Data Mining (DaMi)
Name of the docents	Marco Langenhuizen & Jan Jacobs
Institute	Fontys Hogeschool Techniek en Logistiek
Study Program	Software Engineering & Business Informatics
Year of Study	2017/2018
Location and date of creation	Venlo, December 13, 2017

Contents

Chapter 1

Introduction

1.1 Background

This research has been prepared by two students who are studying Business Informatics & Software Engineering at Fontys University of Applied Sciences, Venlo. The assignment was to run a research for a chosen topic by making use of Machine Learning tools.

1.2 Purpose

The purpose of this research is to analyze peoples drug consumption by looking at people from different countries, with different social states and educational background.

1.3 Demarcation

We do not have the possibility to do a worldwide analysis for this topic because the dataset does not give information on all of it. The main focus is on Australia, Canada, United Kingdom, United States of America, New Zealand and Republic of Ireland. The other countries will be declared with the country "other". Furthermore does our dataset only focus on people who have already made use of drugs once in their life. So people who never made use of any kind of drugs.

1.4 Motivation

In our research we focus on one main question. This question is "Is there any relationship between drug usage and social states? ". To understand the rising usage of drug consumption we are making use of a dataset. This dataset consists of different columns, such as country, age, gender, education, ethnicity, alcohol, caffeine, cannabis, chocolate, coke, ecstasy, heroin, crystal meth and nicotin.

With our research we want to find if there are any relationships between drug usage and social states. For example, do people from a certain country, with a specific educational background, more often make use of a drug like alcohol towards other groups. Besides we want to predict which drug could probably be the next one for you.

Chapter 2

Problem Description

Over the last couple of years the genereal drug consumption is rising. In addition to that, more often people are dying after drug consumption. According to that, in Germany researchers found out, that the number of drug deaths increased by almost 20 percent in 2016, compared to 2014. The following chart only shows the rise in drug consumption in Switzerland, but as you can see the usage of alcohol, nicotine, cannabis, cocaine or amphetamine increased.

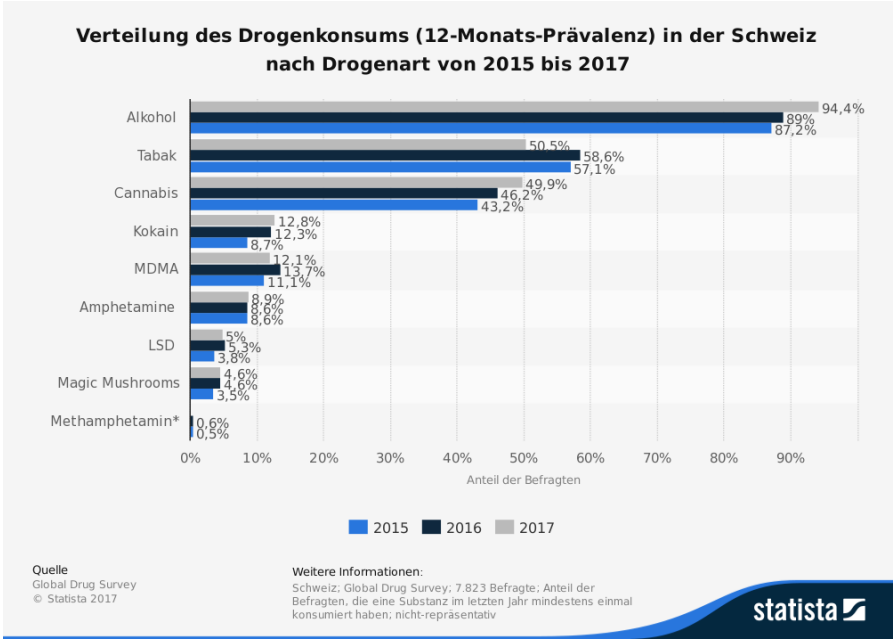

Figure 2.1: Distribution of drug use (12-month prevalence) in Switzerland by drug type from 2015 to 2017 Source: https://de.statista.com/statistik/daten/studie/315194/umfrage/drogenkonsum-in-der-schweiz-nach-drogenart/

Because no-one really knows why the usage is increasing, we are willing to find an answer for that question.

Chapter 3

Research Question

Just by thinking about the mentioned problem in chapter 2, many research questions can be derived. In order to make use of machine learning algorithms, we have to find a question that can be answered by the values which are given in our dataset. Because our dataset consists of values like country, age, education and all different kind of drugs, we came up with the following research question:

Is there any relationship between drug usage and social states?

Drug usage refers to the use of prohibited substances, but also the consumption of commercially available drugs such as alcohol or nicotine. For drug usage, we are taking the values from our dataset for the different kind of drugs into account. They represent the consumption for the drugs like: Never Used, Used over a Decade Ago, Used in Last Decade, Used in Last Year, Used in Last Month, Used in Last Week, Used in Last Day.

The social status refers to the background of the participant. For this purpose, we take a look at the country, together with the educational background of the person. With social states we are talking about the country, where a recipient comes from, but also about the educational background. In our dataset only few countries are represented: Australia, Canada, New Zealand, Republic of Ireland, UK and the USA. If you are from another country you will be represented by the country Other. The educational level is defined as: Low education, Diploma Degree, Masters Degree or Doctors Degree. In our adjusted dataset, the educational level is shown by the numerical values from 0 to 3.

These two values are combined to one value at the end, to have a comprehensive prediction afterwards (e.g. UK-0).

Chapter 4

Path to answer research question

During the research we had a hard start to understand how RapidMiner is working and what we have to do with it. We found out, that we have to format our dataset, because it had some numbers, which were representing differnt values. So at first we formated it in such a way, that it would be understandable for humans, because it is easier to format it in the way, which is needed for the specific algorithm.

For the self-organizing-map (SOM) we separated all countries and made for each country a new column. For all non numerical columns we had to do the same. For RapidMiner we also had to format our dataset in such a way that the Naive Bayes algorithm and the support vector machine can handle the data. This means having a label or do not have any polynomial values.

We decided to use the Naive Bayes algorithm as well as the Support Vector Machine. With the Naive Bayes algorithm we tried to predict the level of education with the information about the usage of different drugs and the age of the person. At first we had to select the attributes, that are needed. For the Naive Bayes algorithm we needed all information about the used drugs and the level of education. We set the role of the education to label, so the model knows with attribute has to be predicted. After training the machine we applied the model and measured the performance of the model.
For the Support Vector Machine we had to make a subset of the attributes. In this case we tried to predict the combination of country and level of education. To use the support vector machine, we had to add a classification by regression function to predict a nominal label using the Support Vector Machine.

Chapter 5

Results

5.1 Machine Learning Tool

To answer our research question, we made use of RapidMiner as Machine Learning Tool because it is "a unified platform to turn data into a strategic asset". (Source: `https://rapidminer.com/products/why-rapidminer/`) Furthermore you can make use of almost every machine learning algorithm, because they are virtually built in. In Addition to that, RapidMiner can be extended to make use of R or Python.

5.2 Naive Bayes

The Naive Bayes algorithm performed with an accuracy of 44%. It is below 50%, so it sounds even worse than random prediction at the beginning, which would be 50%, but it predicts the right value out of 4 possibilities. So a random prediction would be close to 25%. That result tells us that it is not very likely, but in more than 4 out of 10 cases it predicts the right level of education.

accuracy: 43.98% +/- 3.02% (mikro: 43.98%)

	true Diploma Degree	true Doctors Degree	true Masters Degree	true Low Education	class precision
pred. Diploma Degree	192	21	51	162	45.07%
pred. Doctors Degree	3	0	3	0	0.00%
pred. Masters Degree	344	49	173	137	24.61%
pred. Low Education	211	19	56	464	61.87%
class recall	25.60%	0.00%	61.13%	60.81%	

Figure 5.1: Naive Bayes: Prediction of education level

As a conclusion the Naive Bayes algorithm has a relative good accuracy to predict the right value and if it predicts the wrong level of education, it is likely to be just one level above or below.

5.3 Support Vector machine

The Support Vector Machine has an even lower accuracy than the Naive Bayes algorithm. It is around 10%. But in this case it predicts the right value out of 27 different possibilities. In this case a random guess would be close to 3,7%, which is lower than 10%. That means it is not expected, but possible to predict the right value. In general the Support Vector Machine has a pretty low accuracy, but it is still better than a random guess.

5.4 Self-organizing map

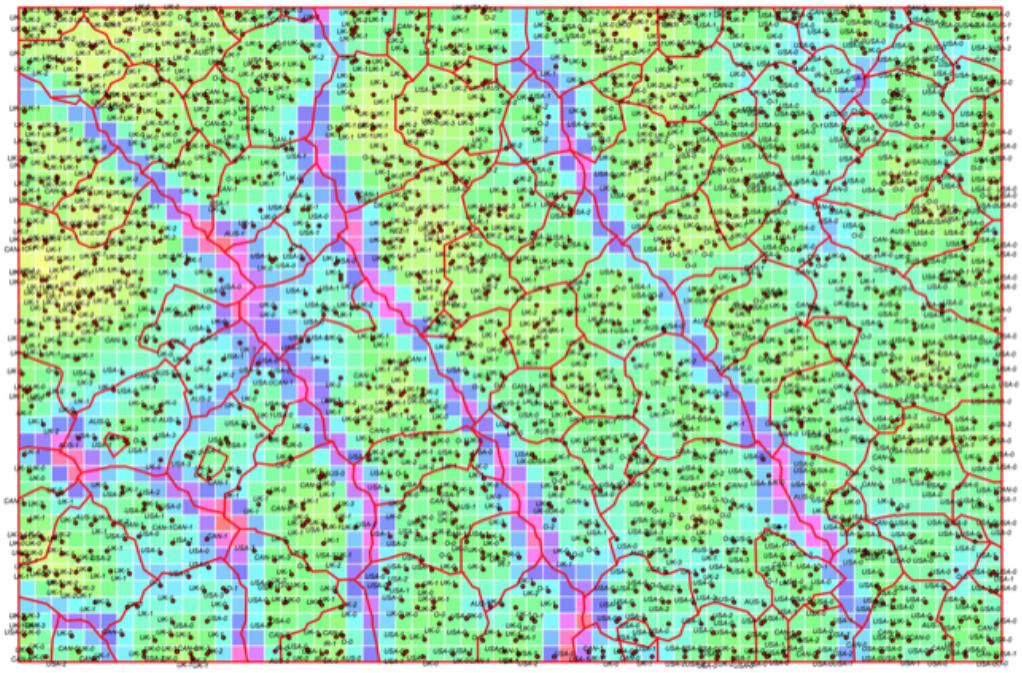

Figure 5.2: Self organizing map: Drug usage

The Self-organizing map (SOM) groups the data pretty good in six different groups of drug-users. That is what we found out on our first look. The next question was, what are these six groups and how do they differ. With a closer look at it, we found out that these groups are the persons at a specific age. Our dataset provided us not with a specific age. It just gave us a range, but the SOM needs a numerical value. So we decided to select for each range one value, which lays in the center of the range. Resulting therefrom our data set already had a grouping by age. The SOM used this grouping and presents it pretty nice.
But this grouping is not useful for our research question. So this SOM was more or less useless for us. At a next step, we could create a new SOM, which does not get the attribute age as a information, so it can not order the values by age anymore. With the new SOM it could be possible to read the different groups as different groups of drug-users, so we creating a new SOM. We decided to keep the attribute age, but we transformed the values in numbers from 1 to 5, which are representing the different groups.

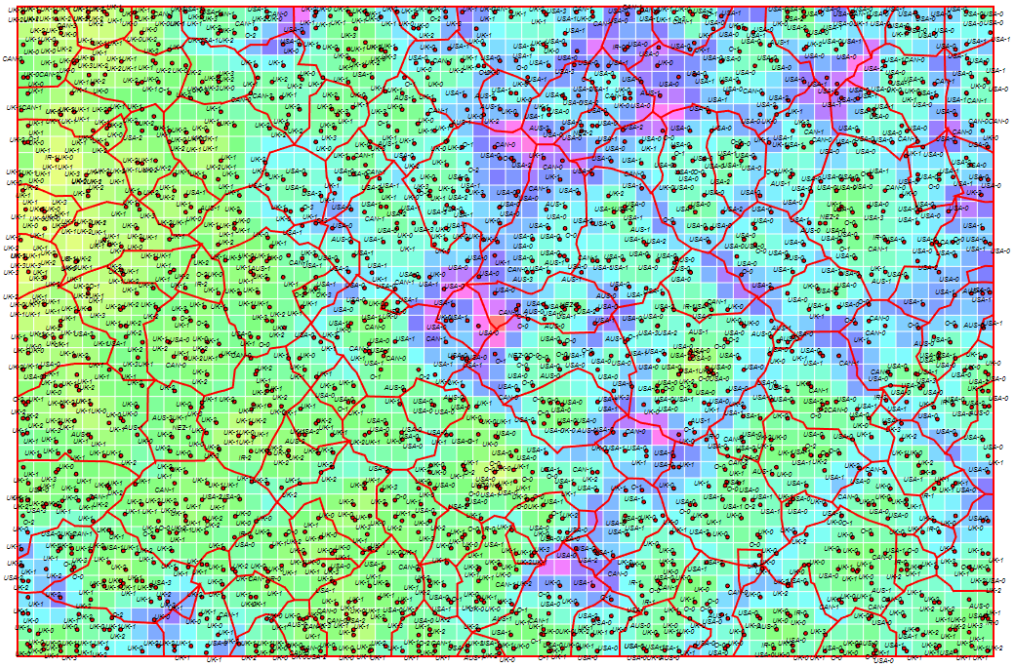

Figure 5.3: Self organizing map: Drug usage with age in groups

The second SOM does not have these clear groups, but it is easy to see that on the left side is a big group and on the right sight are some smaller groups, that are not that clear separated from each other. After a quick look we figured out, that the big group are all the persons, who are taking the "legal" drugs. These are Alcohol, Caffeine and Chocolate. In the bottom left corner, there is a really small group visible. This group represents people, who doesn't drink coffee, but are also taking the other "legal" drugs. On the right side are these kind of people, who are taking a lot of drugs. This group is significant smaller and that is the reason, why there are much more blue areas. For these areas much less or even no data exists. So in general you can say that on the left side are the people, who are taking less drugs and on the right sight are the persons, who are taking a lot of different drugs. By the thicker border in the center, which is separating these two groups, you can say, that some people are more attracted by drugs in general than others. If we look at the other two sides, we cant see a clear separation between the upper and the lower area. These two sides differ in the past of the people. In the upper area are persons, who never tried any drugs. In the lower area are people, who tried some drugs years ago, but did not take them in the last years. This SOM shows us that there is a group of people, that are more attracted by drugs than the others. This could come from their level of education, but can also have other causes.

Chapter 6

Conclusion

After doing the research about drug consumption, we are able to answer our research question, if there is any relationship between drug usage and social states.

After we have used several algorithms, Naive Bayes or Support vector machine, we can state, that there is only a small relationship between drug consumption and social states. How do we come to the conclusion, that there is a small relationship? As shown in chapter 5, we have an accuracy of around 44% with Naive Bayes and almost 10% with Support-Vector machine.

With the help of these two values, we can say that there is a small significant link between drug use and social status, as both percentages are above the value for a random prediction.

To put it in a nutshell, it can be said that the data set was unfortunately not the most suitable one for our research question. There are several reasons for this:

1. not representative enough, because there are only entries about people with drug use
2. not representative enough, as most respondents came from the UK (almost 1000 out of approximately 1900)
3. only six countries were involved. The rest was listed under "other".
4. not representative, since the data record only has an age range and not the exact age

What we can take away from this research is the point that next time the data set should be checked carefully to see whether the possible research question can be answered satisfactorily.