Investigating Tobacco Usage Habits Using Data Mining Approach

Toni Martinović
Faculty of Economics and Business Zagreb, Croatia

Abstract

What are smokers' habits today? Do people rather enjoy cigarettes or rolling tobacco? The research made for this study is going to give us the answer on these questions. The main reason which determines smokers' habits is their lifestyle, e.g. it depends whether they are providing enough money for cigarettes because rolling tobacco is noticeable cheaper. The research is fulfilled by participants of different years, employment status and other lifestyle habits. The research will present the smoking habits of respondents conducted through *data mining*. The data are processed in the *Weka* software with the help of a *decision tree method* - to be precise, the *J48 algorithm*.

Keywords: smokers, habits, tobacco, cigarettes, data mining, WEKA, J48 algorithm, decision trees
JEL classification: D81

Introduction

In many cases the goal of data mining is to induce a predictive model. For example, in business applications such as direct marketing, decision makers are required to choose the action which best maximizes a utility function. Predictive models can help decision makers make the best decision. Supervised methods attempt to discover the relationship between input attributes (sometimes called independent variables) and a target attribute (sometimes referred to as dependent variable). The relationship that is discovered is referred to as a model (Haim Dahan et al., 2014).

The goal of this survey is whether people are more prone to cigarettes or rolling tobacco depending on their lifestyle. The paper starts with basic information about the data mining and some general habits of smokers. The variables that were used in collecting the data, their description, format and modality are presented in the methodology section. This is followed by information regarding the *decision tree* and *J48 algorithm* and principles on which they work. After an introduction to all of the methods of work, a detailed review of the results is shown, with a conclusion based on them. Many different sources were used in the process of making this science paper, such as Witten (2011), Padhye (2006), Palace (1996) and International Journal of Advanced Research in Computer Science and Software Engineering.

Literature review

Data mining

Data Mining is about explaining the past and predicting the future by means of data analysis. Data mining is a multi-disciplinary field which combines statistics, machine learning, artificial intelligence and database technology. The value of data mining applications is often estimated to be very high. Many businesses have stored large amounts of data over years of operation, and data mining is able to extract very

valuable knowledge from this data (Saed Sayad, 2011). It is being used both to increase revenues and to reduce costs. The potential returns are enormous.

Innovative organizations worldwide are already using data mining to locate and appeal to higher-value customers, to reconfigure their product offerings to increase sales, and to minimize losses due to error or fraud. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. The first and simplest analytical step in data mining is to describe the data summarize its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look for potentially meaningful links among variables (such as values that often occur together). As emphasized in the section on the data mining process, collecting, exploring and selecting the right data are critically important. But data description alone cannot provide an action plan. You must build a predictive model based on patterns determined from known results, and then test that model on results outside the original sample. A good model should never be confused with reality (you know a road map isn't a perfect representation of the actual road), but it can be a useful guide to understanding your business. The final step is to empirically verify the model (Two Crows Corporation, 2005).

Tobacco usage habits

Since 1975, the World health organization (WHO, 2015) has stated that smoking tobacco is the single most important factor damaging health to people in the world and to the reduction of smoking made a large, if not most, contribution to the prevention of many diseases. In the civilized world, smoking is spreading from the 15th century. While part of the most developed countries the last decade, putting huge resources into prevention campaigns, managed to some extent reduce the number of smokers, in many underdeveloped and developing countries increased the use of tobacco. Even just a few decades ago smoking was a characteristic of men, and today there are countries where the number of smokers in relation to gender almost equalized. After it began to lose the battle with men's tobacco industry started investing huge funds to promote smoking, focusing their messages towards women, and lately even more to youth. Still it's not easy to explain why so many teenagers begin with cigarette smoking despite knowing about the risks of smoking behavior.

The development of smoking habits can be linked with a number of individual characteristics such as social skills, self-control, self-efficacy and social competence, general personal competence and with a number of other such as the desire to reduce stress, excitement in order to assume the risk, boredom, the need for acceptance of others, the desire for independence and immaturity (Slavko Sakoman et al., 1997).

Methodology

Data description

Below is a table. From Table 1 it can be seen what attributes are set, their descriptions, formats and modalities.

Table 1
Descriptions of attributes, formats and modalities

Attribute name	Description of attributes	Format	Modalities of attributes
Gender	Gender of respondents	Nominal	Female, Man
Age	The age of participants	Nominal	18_25,26_35, 36_and_more
Region	The place where the subjects are	Nominal	Regions
Profession	Occupation of respondents	Nominal	Employed, Student, retired
Are_tobacco_products too_expensive	Opinion about the prices of tobacco products	Nominal	1, 2, 3
Exposure_to_tobacco	Was the subjects exposed to tobacco during growing up	Nominal	1, 2, 3
Impact_of_a_company	To what extent is the company affected a decision about smoking	Nominal	1, 2, 3
Impact_of_courtesy	Did curiosity influenced the decision about smoking	Nominal	1, 2, 3
Sporting	Does respondents do sports	Nominal	Yes, No
How_often_do_you_do_sports	How often respondents do sports	Nominal	1, 2, 3
When_have_you_started_smoking	Says when the subjects started to consume tobacco products	Nominal	Before_18, 18._25, 26._35.
How_much_do_you_smoke	Shows how often the respondents consume tobacco products	Nominal	Every_day, occasionally, weekend
Spending_on_tobacco_products	Speaks about sums of money the respondents spend on tobacco products	Nominal	from_200HRK_to_500HRK,less_than_200HRK,more_than_500HRK
Cigarettes_or_tobacco	The class attribute, includes a selection of types of tobacco products, which subjects consumed, cigarettes or tobacco	Nominal	Cigarettes, Tobacco
Which_cigarettes	Favorite cigarettes brand	Nominal	Marlboro, Ronhil...
Which_tobacco	Favorite tobacco brand	Nominal	Zlatni_Dukat, Bali_Shag..
Will_to_quit_smoking	How much respondents want to stop sm	Nominal	1, 2, 3
Smokers_around_you	How often are the subjects surrounded by smokers	Nominal	1, 2, 3
Tried_to_stop	Speaks about the desire of resp. to stop smoking	Nominal	1, 2, 3
Harmful_impact_of_smoking	Does respondents feel the harmful impact of smoking on their body	Nominal	1, 2, 3

Source: Author's work

Decision trees

A decision tree is a predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes, the branches between the nodes tell us the possible values that these attributes can

have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable. The attribute that is to be predicted is known as the dependent variable, since its value depends upon, or is decided by, the values of all the other attributes. The other attributes, which help in predicting the value of the dependent variable, are known as the independent variables in the dataset. The J48 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained (Neeraj Bhargava et al., 2013).

Results

Decision tree

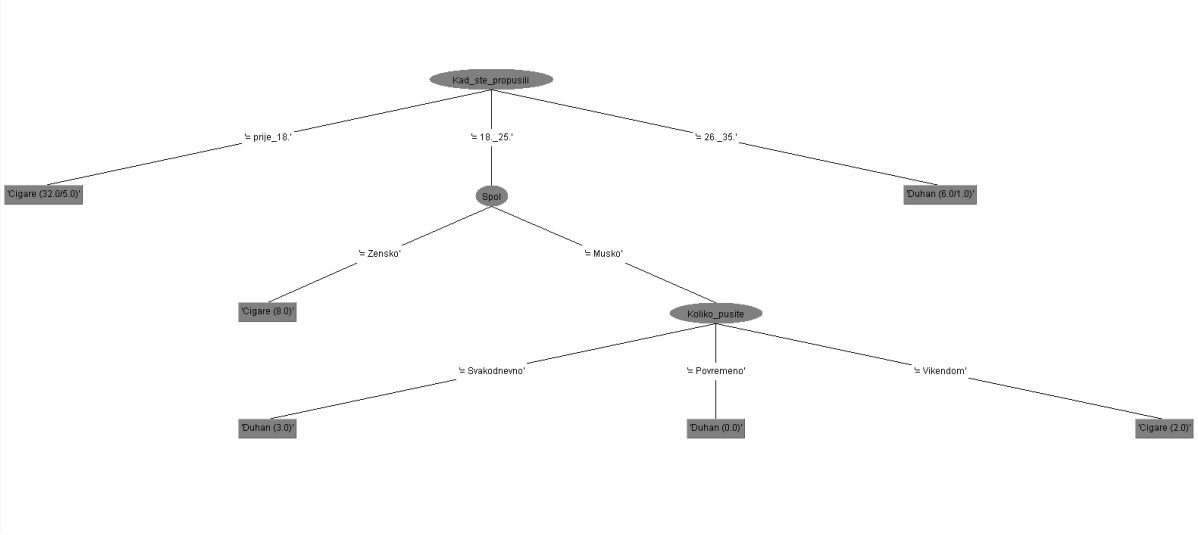
In model development a *J48algorithm* was used, as a model displayed in the form of a decision tree. Number of respondents is 51, and the number of attributes is 8 (How much do you smoke, Gender, When have you started smoking, Spending on tobacco products, Cigarettes or tobacco, Which cigarettes, Which tobacco and Will to quit smoking). *10 fold cross validation* was used to precise the obtained model. Number of trees in the model is 6, and the size of tree is 9.

Leaves of the tree are as follows:

1. If the participants have started smoking before the age of 18, then 27 of them are smoking cigars while 5 of them are wrongly classified.
2. If the respondents started smoking in age of 18 – 25 and if they are female, then 8 of them are smoking cigarettes.
3. If the respondents started smoking in age of18 – 25 and if they are males and they smoke every day, then they smoke rolling tobacco.
4. If everything remains the same, with occasional smoking habit, we have no such examples in the survey.
5. If everything remains the same and if they are weekend male smokers then two of them are smoking cigarettes.
6. If the respondents started smoking between the ages of 26 till 35, then 5 of them are smoking tobacco while one person was wrongly classified.

The habits of smokers can be recognized from the pivot table, i.e. do they rather smoke cigarettes or rolling tobacco, do they smoke before the age of 18. It can be observed that the male and female smokers prefer cigarettes, a total of 52.9% smoke cigarettes (9 of male smokers and 18 female), while only 5 females prefer rolling tobacco. From the respondents between 18 - 25 years, 19.6% of them smoke cigarettes and 3 male subjects consuming rolling tobacco.

Figure 1
Decision Tree



Source: Author's work

In the age range from 26 to 35, only one woman smokes a cigarette and 5 people smoking rolling tobacco. In total, we can say that cigarettes are more desirable than tobacco, i.e. 74.5% smoked cigarettes, and 25.5% tobacco.

Table 2
Pivot table showing smoking habits of females and men toward cigarettes or rolling tobacco

	Before the age of_18.	18._25.	26._35.	Total
Cigarettes	**27**	**10**	**1**	**38**
Male	9	2	1	12
Female	18	8		26
Rolling tobacco	**5**	**3**	**5**	**13**
Male		3	1	4
Female	5		4	9
Grand Total	**32**	**13**	**6**	**51**

Source: Author's work

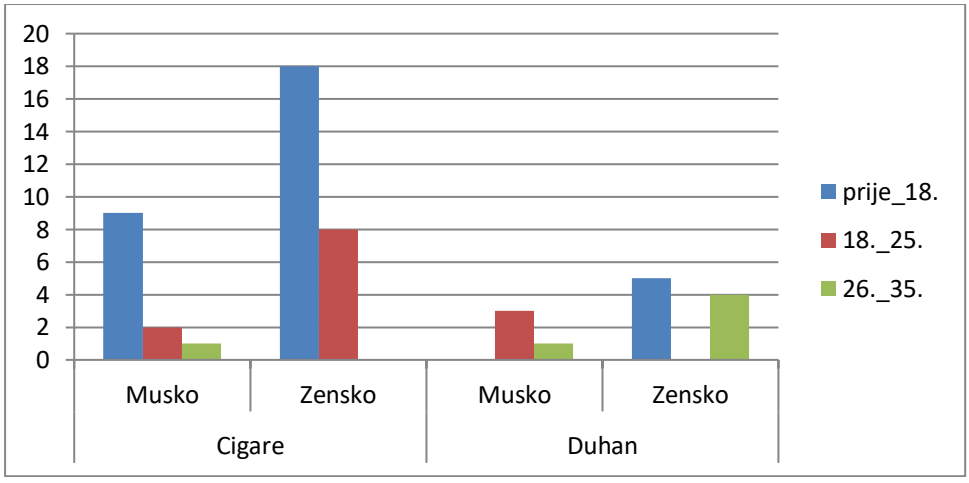
First let's explain a few variables for better result understanding. Kappa Statistic is a measure of the degree of non-random agreement between observers or measurements of the same categorical variable. Mean Absolute Error (MAE) is the average of the difference between predicted and the actual value in all test cases; it is the average prediction error. Relative Absolute Error is the total absolute error made relative to what the error would have been if the prediction simply had been the average of the actual values. (Olaiya Folorunsho, 2013)Of these 51 respondents, 41 of them were accurately classified and the percentage of correctly classified subjects is 80.4%, while 10 respondents were not correctly classified and the percentage of incorrectly classified subjects is 19.6% .Kappa statistic is 0.3914. Mean

absolute error (MAE) is 0.2668. The root mean square error (RMS) is 0.3966 and absolutes Relative Error (RAE) the relative error which is 69.1%.

From confusion matrix can be read out that position AA or TP (True Positive) number is 36, which means that a total of 36 people smoking cigars, while the position BB or IF (true false) shows number of 5 persons who smoke tobacco. The position ba and ab figures are 8 and 2 and they represent wrongly classified subjects.

In the following graph, we can recognize the habits of smokers, i.e. do they rather smoke cigarettes or rolling tobacco, depending on consumption. For respondents who prefer cigarettes and which are men, we see that through all three levels of consumption we have 4 of them. For females who prefer cigarettes, we see that the most represented are those who smokeless than 200 HRK a month, to be precise 17 of them, while the figures reduce as consumption increases. As for the respondents who prefer tobacco, male subjects which consumed 200 to 500 HRK on tobacco products, 3 of them enjoys tobacco. In female subjects, 5 of them spend more than 500 HRK on monthly basis and they are smoking rolling tobacco. In conclusion we can say that those who spend up to 200 HRK monthly prefer cigarettes, while respondents who spend up to more than 500 HRK per month, are smoking rolling tobacco.

Figure 2
The graph shows the habits of respondents to smoking cigarettes or tobacco according to consumption



Source: Author's work

Conclusion

With the help of useful WEKA tools, as tools for analytical processing of data, we were able to create this study to extract information related to smoking decisions of Croats, whether they are more prone to cigarettes or rolling tobacco. The research was conducted through an online survey. The survey respondents had the ability to answer questions about their smoking habits and their decisions about smoking. We received responses from the Croatian regions of Slavonia region, Primorsko goranska county, city of Zagreb and Dalmatia region. The survey was consisted of 20 questions. The first attribute we indicated was gender. Most respondents were females and they counted 35, while the male subjects were 16. Most of the survey respondents were between 18 – 25 years old. The most represented region was Slavonia, while the most represented age group were students (60.8%). The same

percentage of respondents (60.8%) felt that the tobacco products are too expensive, while 47.1% were before heavily exposed to tobacco. 47.1% of respondents expressed that society and curiosity had a big impact on them. All of the respondents were smokers, 70.6% of them were not involved in sports, and the vast majority (70.6%) of them started smoking before the age of 18. 64.7% of respondents were daily smokers and in general spend up to 500HRKon tobacco products on monthly basis. The main tobacco products for most of the respondents were cigarettes (74.5%). Regarding their intent to quit smoking, 39.2% of them wanted to stop, while 49% of them tried to stop more than one time. For the future research, a scale up of questioned respondents is much needed, with a slight change in a survey model. In this model a cumulative effect of respondents which are consuming both types of tobacco products was not questioned, thus making this a future goal. Upon these results a health impact could also be examined, especially regarding the health impact difference between cigarettes and rolling tobacco.

References

1. Bhargava, N., Sharma, G., Bhargava, R., Mathuria, M. (2013), "Decision Tree Analysis on J48 Algorithm for Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3 No. 6, pp. 1114-1119.
2. Dahan, H., Cohen, S., Rokach, L., Maimon, O. (2014), "Proactive data mining with decision trees", Springer-Verlag, New York.
3. Folorunsho, O. (2013), „Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3 No. 3, pp. 11-15.
4. Padhye, A. (2006), "Natural Language Processing", Duluth: University of Minnesota, USA.
5. Palace, B. (1996), "What is data mining", Los Angeles: Anderson Graduate School of Management at UCLA, USA.
6. Sakoman, S., Kuzman, M., Šakić, V. (1997), "Pušačke navike zagrebačkih srednjoškolaca" [Smoking Habits of the High School Pupils in Zagreb], Društ. Istraž. Zagreb, Vol. 30-31 No. 4, pp. 513-535.
7. Sayad, S. (2011), „Real time data mining", Cambridge Ontario: Self-Help Publishers, USA.
8. Two Crows Corporation (2005), „Introduction to Data mining and knowledge discovery", 3rd ed., USA.
9. Witten, I.H. (2011), "Data mining: Practical machine learning tools and techniques", Burlington: Morgan Kaufmann Publishers, USA.
10. WHO (2015), World Health Organization, available at: http://www.who.int (accessed June 26th 2015).

About the author

Toni Martinović is currently a master student at the Faculty of Economics and Business, University of Zagreb, where he is enrolled into the Master study of Managerial Informatics. He earned his B.Sc. of Management at the Faculty of Economics and Business, University of Rijeka. Currently he is doing his master paper at the department of Informatics, in the field of Managing Innovations. His hobbies include gym, handball playing, playing guitar and animal welfare. Author can be contacted at **tonimartinovic11@gmail.com**